FLOOD FORECASTING ON THE HUMBER RIVER USING AN ARTIFICIAL NEURAL NETWORK APPROACH

HAIJIE CAI

FLOOD FORECASTING ON THE HUMBER RIVER USING AN ARTIFICIAL NEURAL NETWORK APPROACH

by

Haijie Cai, B. Eng, M. ASc

A thesis submitted to the School of Graduate Studies

in partial fulfillment of the requirements for the

degree of Master of Engineering

Faculty of Engineering and Applied Science

Memorial University of Newfoundland

2010

St. John's Newfoundland Canada

ABSTRACT

In order to provide flood warnings to the residents living along the various sections of the Humber River Basin, the Water Resources Management Division (WRMD) of Department of Environment and Conservation, Government of Newfoundland and Labrador has generated flow forecasts for this basin over the years by means of several rainfall-runoff models. The first model used is the well-known Streamflow Synthesis and Reservoir Regulation Model (SSARR) which is a deterministic model that accounts for some or all of the hydrologic factors responsible for runoff in the basin. However, the accuracy of the model became worse over the years. Although it was calibrated well in the beginning, recalibration of the model has not been very successful. In addition, the model cannot take into account the snowmelt effect from the Upper Humber basin. The next model is the Dynamic Regression model, a statistically based model that uses the time series of historic flows and climate data of the basin to generate a forecast. This model was tried during the late 1990s to early 2000s. This model was found to provide better forecasts than the SSARR model, but it also does not take into account the snowmelt effect from the upper regions of the Humber River. The third model tried by the WRMD was an in-house Routing model. This method uses a series of water balance equations which can be easily implemented on a spread sheet at each gauging station. However, calibration is done subjectively and the forecast obtained for the snowy region of the Upper Humber is still a problem. In view of the foregoing issues with the above models, a better model that is easy to use and calibrate, provides accurate forecasts, and one that can take into account the snowmelt effects is required. Since 2008, the WRMD has been using the statistically based Dynamic Regression Model on an interim basis until a replacement model could be developed.

This thesis presents the development of artificial neural network (ANN) models for river flow forecasting for the Humber River Basin. Two types of ANN were considered, general regression neural network (GRNN) and the back propagation neural network (BPNN). GRNN is a nonparametric method with no training parameters to be adjusted during the training process. BPNN on the other hand has several parameters such as the learning rate, momentum, and calibration interval, which can be adjusted during the training to improve the model. A design of experiment (DOE) approach is used to study the effects of the various inputs and network parameters at various stages of the network development to obtain an optimal model. One day ahead forecasts were obtained from the two ANNs using air temperature, precipitation, cumulative degree-days, and flow data all suitably lagged (i.e. of 1 day or 2 day before) as inputs. It was found that the GRNN model produced slightly better forecasts than the BPNN for the Upper Humber and both models performed equally well for the Lower Humber. The ANN approach also produced much better forecasts than the routing model developed by the WRMD but was not much better than the dynamic regression model except for the Upper Humber.

ACKNOWLEDGEMENTS

I would like to extend my sincere thanks to my supervisor, Dr. Leonard Lye for all his academic guidance, thesis supervision, and financial support, throughout my program. My appreciation is also extended to the Newfoundland and Labrador's Department of Environment and Conservation, especially to Dr. Amir Ali Khan for providing advice and access to the data used in this thesis. I would also like to thank the Institute of Biodiversity, Ecosystem Science and Environmental Sustainability (IBES) for providing additional funding for this study and the Faculty of Engineering and Applied Science for their general support and providing teaching assistantships during my studies at MUN. Last but not least, I would like to thank my parents for their continuing encouragement and support during the last few years.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

WRMD	Water Resources Management Division
ANN	Artificial Neural Network
BPNN	Backpropagation Neural Network
GRNN	General Regression Neural Network
SSARR	Streamflow Synthesis and Reservoir Regulation Model
DOE	Design of Experiment
RSM	Response Surface Methodology
CCD	Central Composite Design
SHE	Systeme Hydrologique Europeen Model
IHDM	Institute of Hydrology Distributed Model
DD	Cumulative Degree Days
QB	Flow at Black Brook
QR	Flow at Reidville
QV	Flow at Village Bridge
TB	Air temperature at Black Brook
TA	Air temperature at Adies Lake
PB	Precipitation at Black Brook
PA	Precipitation at Adies Lake
WL	Water Level of Deer Lake
DF	Degrees of Freedom

Chapter 1

Introduction

1.1 Background

In recent decades, studies on rainfall-runoff relationships have become more and more important to various communities along rivers because of the increase in development and subsequent damage of floods. Traditionally, rainfall-runoff relationships are studied using models that are based on a collection of principles set out in mathematical form that attempt to describe the characteristics of a river basin. These mathematically based hydrologic models are normally called conceptual rainfall-runoff models. They have several physical parameters such as drainage area and stream slope, and process parameters such as depths of the water table, interflow rates, coefficients of infiltration, percolation and soil storage that need to be defined along with the precipitation inputs. Many such models are used in hydrology for various purposes. Some of these models include the Streamflow Synthesis and Reservoir Regulation (SSARR) Model, Systeme Hydrologique European (SHE) Model, Institute of Hydrology Distributed Model (IHDM), Kinematic Wave Model, and many others (Beven 2001). For flood forecasting purposes, some of the available models include the Lambert ISO Model, and TOPMODEL, and many other statistically based models (Beven 2001). Many of these models are site specific and the success or failure of the chosen model is usually dependent on the extent and quality of data available.

For flow forecasting on the Humber River Basin, the Water Resources Management Division (WRMD) of the Department of Environment and Conservation, Government of Newfoundland and Labrador has used several models over the last 20 years. The deterministic SSARR model

was first applied during the late 1980s (Cummin and Cockburn 1986). A newer version of the SSARR Model was later developed and applied by the WRMD which used only the daily average temperature and total daily precipitation (Picco 1996). The results of the SSARR model underestimated the runoff from the rainfall/snowmelt event in December and overestimated the snowmelt in April, which indicated that the model did not simulate enough snow melt in November/December. Since the Upper Humber is mostly covered by snow during the winter, the poor performance of snowmelt simulation of the SSARR model affected the accuracy of the forecasts. In addition, there was difficulty with the recalibration of the model, lack of technical support for the software, and the software is both cumbersome to use and is now practically obsolete.

The next model to be tried by the WRMD was a statistically based Dynamic Regression Model (Picco 1996). This is a linear time series model where the flow forecasts were generated using lagged flows and precipitation as inputs. Dynamic regression models were developed at various flow gauging stations along the Humber River. While this approach provides better forecasts than the SSARR model, it also cannot take into account the snowmelt effect from the upper regions of the Humber River. In addition, as the model used a simple linear regression approach, any nonlinear hydrologic effects could not be captured by the model.

The next model used by the WRMD was an in-house Routing model. This method uses a series of water balance equations which can be easily implemented on a spread sheet at each gauging station (Rollings 2008). However, calibration is done subjectively with model parameters determined on a trial and error basis. The drawback of the model is that it is not able to

incorporate the snowmelt from snow covered region of the Upper Humber effectively. Since 2008, WRMD has been using the statistically based Dynamic Regression Model on an interim basis until a replacement model can be developed.

In view of the foregoing issues, the WRMD seeks a better model that is easy to use and calibrate, provides accurate forecasts, and can easily provide snowmelt simulations for the Upper Humber basin. In this thesis, a non-conceptual flow forecast model based on an artificial neural network (ANN) is proposed for this basin. ANN is a relatively new methodology that has been used in many areas other than hydrology. This model will be discussed in detail later and the developed model will be tested against the currently used models during the 2009 flood season.

1.2 Description of the Study Area

The Humber River Basin is located on the west side of the Island of Newfoundland in Canada. The total length of the Humber River is about 153 km. The head waters are located in the Long Range Mountains (elevation around 800 m) the north western side of the Island. The drainage area is over 8,000 km^2 which makes it the second largest basin on the Island. The basin has a humid continental climate with temperatures ranging from about -25°C to 20°C. The whole Humber River Basin can be divided into two main parts by Deer Lake (Figure 1.1). The Upper Humber is in the northern part of the basin, most of which is located in the mountainous area. The elevation of the Upper Humber River around Black Brook is between 600m and 800 m. This region is normally covered by snow during the whole winter from October to April. The stream flow at Black Brook during the spring is thus strongly influenced by snowmelt. The Lower Humber represents the southern part of the basin which contains the plains of Deer Lake and

Grand Lake. The average elevation of the Lower Humber is about 100 m. Most of the lower basin is regulated for hydroelectric power generation by the Deer Lake Power company.

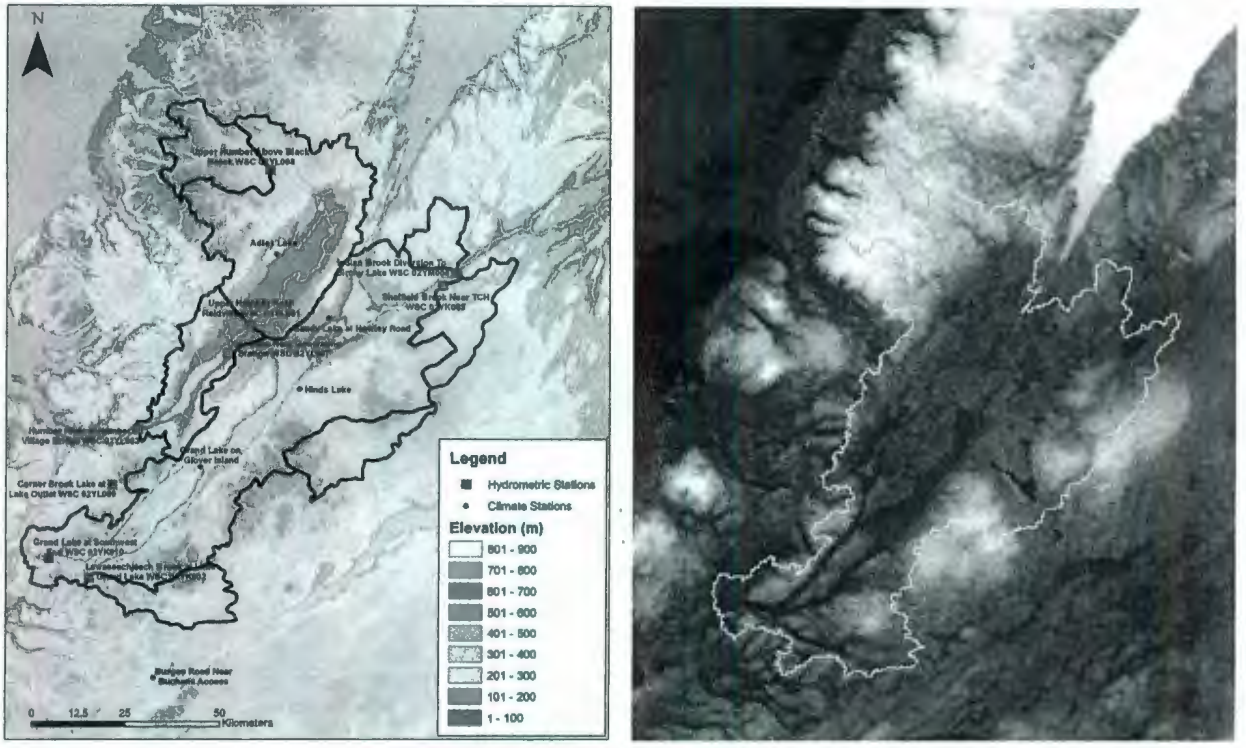


Figure 1.1 Map and satellite image of the Humber River Basin (Water Resources Division, Department of Environment and Conservation, 2007)

Over the years more and more people have come to and live along the Humber River at Humber Village Bridge close to Deer Lake and Steady Brook. The current population has already grown to over 25,000 (Statistic Canada 2006). During the development of the communities many houses were built in the flood plains that are frequently subjected to floods during the spring flood season. In order to protect the communities from flood damage, an accurate and timely forecast of the flow of Humber River is necessary so that the residents living along the Humber River can be warned and preventative action taken ahead of any impending floods (Picco 1996).

1.3 Data Available

As shown in Figure 1.1, there are 8 hydrometric stations and 8 climate stations in operation both around and within the Humber River basin. Concurrent daily records for various climatic and flow variables are available from 1997 to 2008 although flow data are available at some stations from 1920 (Figure 1.2). The data for this study were provided by WRMD.

Although the data available were from 1997 onwards, there is an unfixable problem with the temperature data from 2001 to 2002 at Black Brook. Because of this, all concurrent data during that period are not used. Therefore, only 9 complete years of data are available at Black Brook. For the other two stations, Reidville and Village Bridge, the data from 1999 to the middle of 2008 are all available. As three individual models will be developed for each of the three stations, the data are arranged differently according to the requirement of the models. The Reidville station does not have climate data. Therefore climate data at Adies Lake will be used for the model at Reidville because of the proximity of Adies Lake to Reidville and they are on the same plain. For the model at Village Bridge, only hydrometric data are used since there is no nearby climate station. Flow records are also available at other stations on the Humber River at Grand Lake outlet since the early 1920s and at Village Bridge since the early 1980s. But there were insufficient concurrent climate data available for model development.

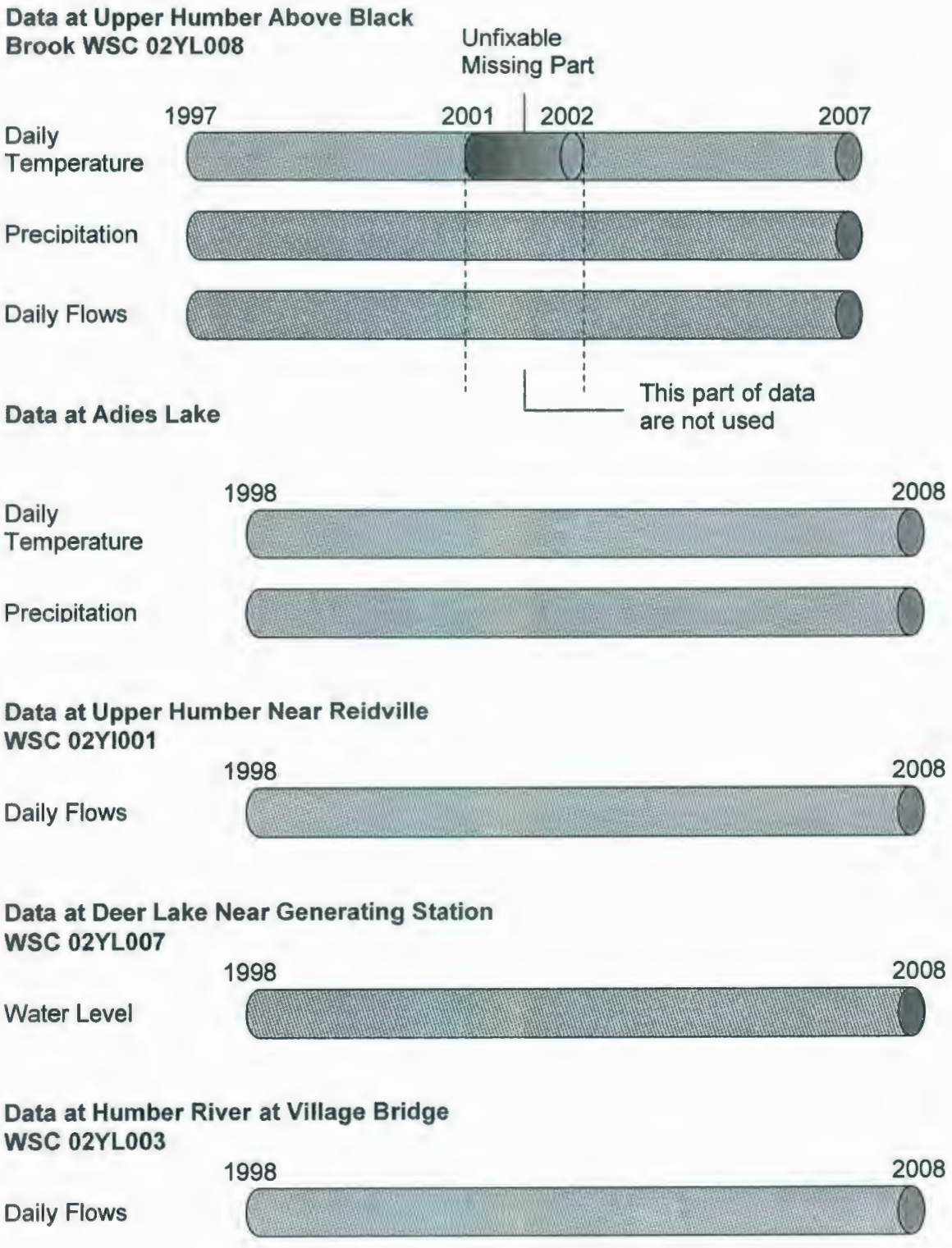


Figure 1.2 Data availability at each station

1.4 Study Objectives

The objectives of this study are threefold:

1. To investigate the use of various ANN models for real-time flood forecasting along the Humber River. In particular, the back propagation neural network (BPNN) and the general regression neural network (GRNN) models will be investigated.
2. To investigate the use of Design of Experiment (DOE) methodology for the calibration and input selection of ANN models. A two-level factorial design will be used to investigate the sensitivity of the various input factors on the accuracy of forecasts, and to calibrate the ANN models to obtain an optimal ANN.
3. To validate the results from the developed ANN models and compare with the currently used models. The forecasts obtained from the two types of ANN models will be compared to the Dynamic Regression and routing models currently used by the WRMD for data collected during the 2008-2009 flood season.

1.5 Outline of the Thesis

The background for the thesis, the description of the study area, and objectives of this thesis have been presented in the previous sections. The general idea of rainfall-runoff modelling and the models which were used on the Humber River Basin are discussed in Chapter 2. The theoretical considerations of Artificial Neural Network (ANN) are described in Chapter 3. The methodology used for calibration and verification of the ANN is discussed in Chapter 4. The results obtained from the ANN models at the three gauges and how they compare to the currently used models are presented in Chapter 5. Finally, the conclusions and recommendations of the study are presented in Chapter 6, followed by the references.

Chapter 2

Hydrologic Modelling

As mentioned in the last chapter, the WRMD has used several models for flow forecasting on the Humber River Basin. In this chapter, a brief overview of hydrologic modelling is given followed by brief reviews of the three models used by WRMD. These are the deterministic SSARR model, the statistically based Dynamic Regression Model and the in-house Routing model.

2.1 General Principles of Hydrologic Modelling

Precipitation, runoff, and evaporation are the principal processes that carry moisture from one system to another. When the moisture of the earth system is considered, three systems can be distinguished: 1) the land system, 2) the subsurface system, and 3) the aquifer system. Streamflow in a perennial river is derived from these systems. In the land system, precipitation, surface runoff, infiltration, and evapotranspiration are the dominant processes generating and abstracting moisture. When physiographical and structural characteristics of different locations are considered, interception, depression, and detention storage are also used to describe the moisture movement. The moisture can be lost to the atmospheric system or subsurface system through these processes. Hydrologic models aim to quantify and model all these processes that govern moisture through the various systems. In this regard, the rainfall-runoff relationship is one typical process of moisture movement and modelling of this relationship is valuable in many aspects (Singh 1989).

2.2 Rainfall-runoff modelling

Since hydrological measurement techniques are limited, it is not always possible to measure everything necessary to understand a particular hydrological system. In fact, aside from the limited range of techniques, there are also limits on space, and time for measurements. To extrapolate the information that is limited it is necessary to take advantage of the measurements that are available. Rainfall-runoff modelling is one such tool that can be used to provide the means to quantitatively extrapolate or predict hydrologic response which is helpful in decision making concerning a particular hydrological problem.

Rainfall-runoff modelling can be carried out using two main approaches. In the first approach, the models can be described by some physical interpretations based on an understanding of the nature of catchment response. This approach is generally data intensive and has many parameters that require calibration. For the second approach, the models are based purely on an analytical framework which uses only observations of the inputs and outputs to a catchment area. Model parameters are fewer and they are estimated using observed data. This approach to modelling can be described as a 'black box' approach, which does not refer to the internal processes that control the rainfall to runoff transformation. Both approaches to rainfall-runoff modelling however do require some understanding of the catchment processes and the availability of suitable data.

Rainfall-runoff models rely heavily on rainfall records. These records are measured by point rain gauges in monthly, daily or shorter time steps. In large catchments, daily time step may be sufficient for practical modelling purposes. The spatial variation of inputs in large catchments is generally more important than the temporal variation. In smaller catchments, the daily time step

may be longer than the storm response time of the catchment hence a finer time resolution may be required for accurate modelling of the rainfall-runoff response. The accurate measurement of rainfall is very important in rainfall-runoff modelling to produce accurate runoff predictions. No model will be able to give good predictions if the inputs to the model do not adequately characterize the rainfall inputs from the catchment.

Runoff generation controls how much water gets into the stream and flows towards the catchment outlet. The runoff not only takes into account of the rainfall intensity during the time-frame of storm but also considers the routing of the runoff from the source areas to the outlet. The routing only depends on the flow processes within the stream, which can be reasonably well described on the basis of hydraulic principles. Therefore, every rainfall-runoff model requires two essential components. One is to determine how much of a rainfall become part of the storm hydrograph; the other is to take into account the distribution of that runoff to form the shape of the storm hydrograph.

In addition to the rainfall and runoff information, some other input variables also play important roles in the model. These include evaporation, interception, snowmelt, and catchment physical characteristics. In many environments, especially Canada, snowmelt may be the most important source of the annual maximum discharge in most years and may be a major cause of flooding. The processes of snow accumulation and rate of melt are therefore also required in the rainfall-runoff model. The data requirements for different snow models are varied depend on the snowmelt method. The most common and simple approach is the temperature index or 'degree-day' method. This method is based on the hypothesis that snowmelt is proportional to the

difference between air temperature and a threshold melt-temperature. The degree-day method is simple and it has the advantage of demanding only temperatures as an input. In addition, this method gives good performance when snow melt is dominated by heat input due to radiation (Beven 2001).

2.3 Rainfall-runoff models used in the Humber River Basin

In the next sections, the three models that have been tried over the years by WRMD for flow forecasting in the Humber River will be briefly described. These are the SSARR, Dynamic Regression, and Routing models.

2.3.1 Streamflow Synthesis and Reservoir Regulation Model (SSARR)

The SSARR Model was first developed in 1956 by the U.S. Corps of Engineers (North Pacific Division) for planning, design and operation of water control works. It was further developed for operational river forecasting, river management activities, and reservoir regulation for several major projects on large rivers such as Columbia River and Mekong River. Both rainfall and snowfall events are considered in the model. This model has more than 24 parameters and some of them are lumped. These parameters are usually adjusted by trial-and-error optimization. The inputs to this model include: daily rainfall, daily temperature, insolation, and snowline elevation. The output is the daily streamflow. The interval of calculation can be from 0.1 to 24h. This model is much more simplified in representation of catchment components than other models such as the Stanford Watershed Model (Singh 1989).

The SSARR model can be described as a closed hydrologic system in which the water budgets are defined by meteorological inputs (rainfall and snowmelt) and hydrologic outputs are defined by runoff, soil storage and evapotranspiration losses. As a deterministic hydrologic watershed model, some underlying principles must be preserved. Firstly, the basic elements in the hydrologic cycle such as rainfall, snowmelt, interception, soil moisture, interflow, groundwater recharge, evapotraspiration, and the various time delay processes should be accounted for while processing the objective function that relates them to observed hydrometeorological variables. The level of complexity that the model uses to represent a particular process depends on what elements are selected.

Second, the SSARR model contains streamflow routing functions which provide a generalized system to solve the unsteady flow conditions in river channels where streamflow and channel storage effects are related, either at one point or a series of points along a river system. The streamflow routing functions can be applied in many ways depending upon the type of basic data available, and the conditions of the river system with respect to back water effects from variable stage discharge effects, such as tidal fluctuations or reservoir fluctuations.

Third, the SSARR model was designed to include the effects of reservoirs or other water control elements within the streamflow simulation process. Reservoirs may be described for any location in the river system, while inflows are defined from single or multiple tributaries. These inflows can be derived either from watershed simulation for river basin upstream or from specified flows as a time series, or a combination of the two.

Last, the outflows from reservoirs are determined based on all the principles above. In order to provide a once-through process for the system as a whole, the basic hydrologic elements, channel storage effects, and reservoirs or other water control elements need to be considered sequentially in the river basin simulation.

One of the useful features of the SSARR model is that it allows the distribution of data from a number of meteorological stations to the subbasins defined in the model. This advantage is particularly helpful in representing the hydrologic regimes of large basins, like the Humber River basin, because it accounts for the spatial variations of the meteorological parameters.

Cummin and Cockburn (1986) had developed the SSARR model in 1984 and 1985 for the Humber River Basin to assess the possibility of using the SSARR model to forecast flows on the Humber River during high flow events. The study found that the data collection network needed to be improved if the model was to produce accurate flow forecasts. Additional stations with temperature and precipitation sensors and transmitters were thus installed for near real time data acquisition.

A newer version of the SSARR Model was later developed and applied by the WRMD which used only the daily average temperature and total daily precipitation (Picco 1996). The main elements of the Humber River watershed are represented in the SSARR model by 11 sub-basins, two reservoirs and one lake. The meteorological data can be weighted from each station for each basin by the user. In this process, the temperature and precipitation data were first given the same weights for each station by Picco (1996). Then the weights were adjusted during the calibration

process. The snowmelt coefficients and routing coefficients are two additional parameters that were then calibrated. The snowmelt coefficient accounts for the estimation of rain freeze temperature, base temperature, lapse rate, and melt rate. Routing coefficients were determined for the runoff conditions which were independent of snowmelt. The precipitation and temperature weighting coefficients, snowmelt coefficients, and routing coefficients are given in Picco (1996).

The information was saved in punch cards when the SSARR model was first developed in the 1950's. The cards were marked by a specific code for different types of data. For example, precipitation data are stored on a "Z4" card. In each card, the station information, time period of the data and actual data are identified. The original SSARR model has now been converted from a mainframe computer to microcomputer use. However, the format of the data is still stored in card format as separate data files. All the input and output information are then acquired from the "card file". Due to the obsolescence and lack of technical support, the model is currently difficulty to recalibrate.

When the SSARR model was applied in Picco (1996), the results from the model underestimated the runoff from the rainfall/snowmelt event in December and overestimated the snowmelt in April, which indicated that the model did not melt enough of the November/December snow. Since the Upper Humber is mostly covered by snow during the winter, the SSARR model was not able to adequately take into account the snowmelt effect from the upper Humber basin which affected the accuracy of the forecasts.

2.3.2 Dynamic Regression Models

A Dynamic Regression model is a kind of single equation regression model which combines the time series features with the effects of explanatory variables. The special feature of time series is that the output variables are correlated through time rather than being independent. However, a dynamic regression model will also consider the influences of explanatory variables in addition to the time series propagation.

There are two preconditions when we intend to use dynamic regression models:

1. Enough and stable data to support a correlational model, because many time series such as temperature readings and river flows exhibit annual variation. For example, temperature is high in summer and lower in winter.
2. The additional explanatory variables such as daily average temperature, daily total precipitation increase the performance of the model in a meaningful way. Otherwise, a purely dynamic model would be sufficient.

Figure 2.1 shows the procedure used to develop the dynamic regression models.

Dynamic Regression Model Building Cycle

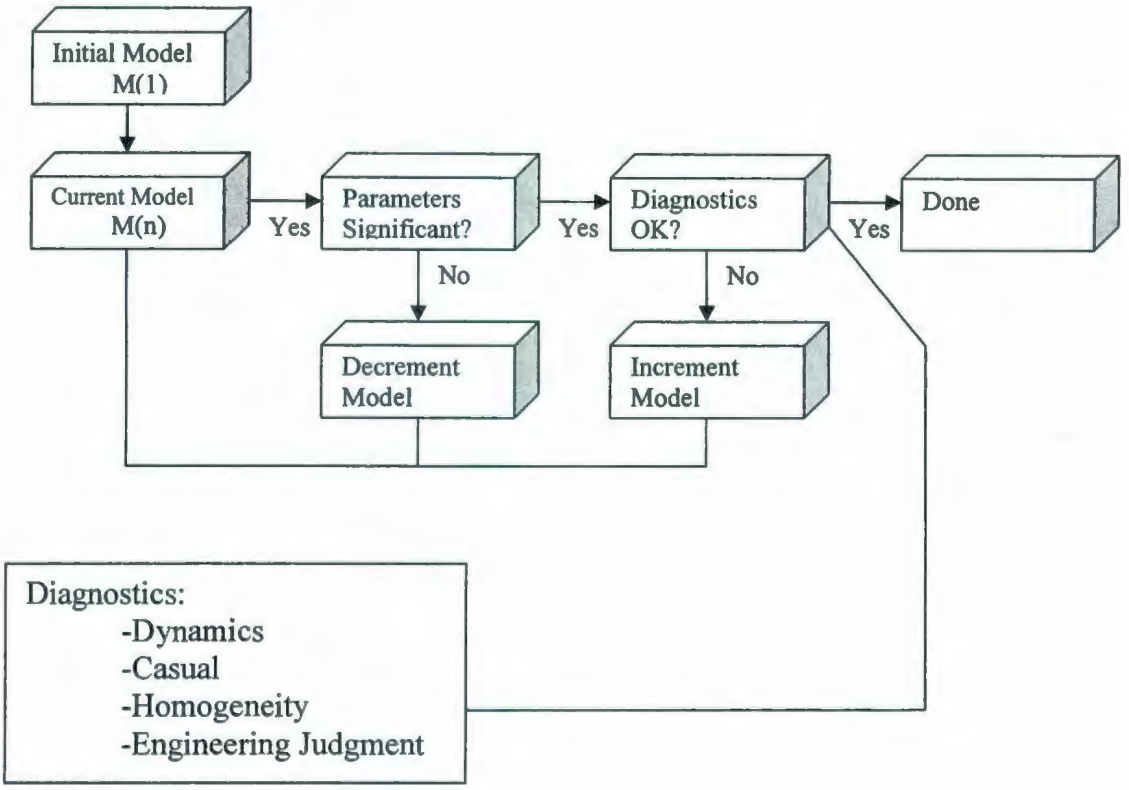


Figure 2.1 Dynamic Regression Model Building Cycle (Goodrich 1989)

The procedure starts with the simplest form of the regression relationship and then builds on that relationship until the best fit to the data is obtained. The parameter's significance test is used to determine if the variable is important for the model. When all the variables that are statistically significant are in the model, the diagnostics tests for the model are run. This part of the model building process mainly focuses on the lagged variables and autoregressive terms. During this phase, some new lagged variables or autoregressive terms may be introduced to the model as some gauges are linked to each other. The Forecast Pro software package can provide a calculation of various goodness of fit tests: mean absolute deviation, standard forecast error, r-square value, Bayesian information criterion, Durbin-Watson test, and the Ljung-Box test. This

whole procedure is continued until a satisfactory result is achieved. The model at each gauge can be expressed as:

FlowB[t] = Constant + a Precipitation + b FlowA[t] + c Precipitation[t-1] + d FlowA[t-1] + e FlowB [t-1] + ... + AUTO [ ] (Eq 2.1)

Where: Flow A and Flow B are linked to each other;

[t-1] is the value of 1 day before, that could be [t-2], [t-3], and so on

a, b, c, d, e, f, ...are the coefficients calculated for each significant variable;

AUTO[] is the autoregressive error term.

Goodrich (1989) used a Cochrane-Orcutt model to improve the model dynamics by introducing new parameters. With his method, Eq 2.1 is replaced by:

$$\phi(b)Y_t = \beta Z_t + \omega_t \qquad \text{(Eq 2.2)}$$

$$R(b)\omega_t = \varepsilon_t \qquad \text{(Eq 2.3)}$$

Where $\phi(b)$ = autoregressive polynomial;

Y_t = dependant variable at time t;

β = coefficient of i th exogenous variable $Z_t^{(i)}$;

Z_t = vector of exogenous variables at time t;

$R(b)$ = polynomial in the backward shift operator;

ω_t = raw residual at time t; and

ε_t = errors where the errors are $NID(0, \sigma^2)$, ie. normally and independently distributed with variance σ^2.

Eq 2.2 and Eq 2.3 can also be written in a single equation as:

$$R(b)(\phi(b)Y_t - \beta Z_t) = \varepsilon_t \qquad \text{(Eq 2.4)}$$

The "dynamic regression" model was used in 1991 by (Pankratz 1991) who referred to a technique called "combined transfer function-disturbance" by Box and Jenkins in 1976. The ordinary least squares dynamic regression model takes the form:

$$\phi(b)Y_t = \beta Z_t + \varepsilon_t \qquad \text{(Eq 2.5)}$$

Where $\phi(b)$ = autoregressive polynomial

Y_t = dependant variable at time t;

β = coefficient of ith exogenous variable $Z_t^{(i)}$;

Z_t = vector of exogenous variables at time t, i.e. temperature or precipitation; and

ε_t = errors where the errors are $NID(0,\sigma^2)$, ie. normally and independently distributed with variance σ^2.

Usually, the residuals from Eq 2.1 are correlated, contrary to the assumption of independence. This significant correlation indicates that the historical data are related to current data or future values. In order to estimate the autocorrelations, the autocorrelation function can be tested by the Ljung-Box Q-test, Durbin-Watson test or any other tests can be used. The autocorrelation function can determine if one or more lags should be added to the model or additional exogenous variables such as temperature or precipitation should be added.

Picco (1996) used dynamic regression models to develop forecasts on Humber River Basin in 1996 and 1997. The procedure was carried out using the Forecast Pro Software package. Hydrometric data and climate data were used for each gauging station when data are available. The climate data closest to the gauge were used if there are no local climate data available. Usually flood forecast is performed from April to June and September to December. There is no chance of flooding in January and February due to subzero temperature. The table below shows the various equations used to run the Model.

Sub-basin Name	Form of Dynamic Regression Equation
Lewaseechjeech Brook	_ CONST + a PREGLGI + b FLOW[-1] + c FLOW[-2] + d FLOW[-3]; where: _CONST = 0.149490 a = 0.244776; b = 1.664595; c = - 1.046278 d = 0.336051
Sheffield Brook	_ CONST + a PREINDI + b FLOW[-1] + c FLO W[-2]; where: _ CONST = 0.348126 a = 0.041432; b = 1.432156 c = -0.462627
Indian Brook Diversion	_ CONST + a PRECINDI + b FLOW[-1] + c FLOW [-2]; where _ CONST = 0.448615 a = 0.118321 ; b = 1.297039; c = -0.362822
Upper Humber River above Reidville	_ CONST+ a PRESAND + b FLOW[-1] + c FLOW[-2] + d FLOBLAC + e _AUTO[-1]; where: _ CONST =14.380586 a = -0.210896; b = 0.739059; c = -0.376750 d = 1.055177; e = 0.884608
Upper Humber River above Black Brook	_ CONST + a PRECBLAC + b FLOW[-1] + c FLO W[-2]; where: _ CONST = 1.254866 a = 0.558944; b = 1.238943; c = -0.315646
Humber River at Humber Village Bridge	_ CONST +a PREBLAC + b FLOW[-1] + c FLOREID + d FLOBLAC + e _AUTO[-1]; where: _ CONST = 21.697851 a=0.187210; b =0.859492; c =0.196181 d = -0.055710; e = 0.525336

Table 2.1 Various forms of Dynamic Regression Models

For example, the model for calculating the flow of the Upper Humber River at Reidville generated by the dynamic regression method is shown in Eq 2.6:

FLOWREID [t] = 14.380586 - 0.210896 PRESAND[t] + 0.739059 FLOWREID[t-1] – 0.376750 FLOWREID[t-2] + 1.055177 FLOBLAC[t] + 0.884608 AUTO[t-1] (Eq 2.6)

Where: FLOWREID means the flow at the Reidville station, PRESAND means the precipitation at Sandy Lake, and FLOBLACK means the flow at Black Brook

This equation shows that the flow of the Upper Humber River at Reidville at Day t is related to the precipitation at Sandy Lake at Day t (which is close to the gauge at Reidville), the flow at Reidville of 1-day and 2-days before, and the flow at Black Brook at Day t. Among them, the flow at Black Brook plays the most important role in this model since it was with the highest coefficient. Therefore, the flow at Reidville is not the only time series involved but it highly depends on the flow at Black Brook and precipitation at Sandy Lake. The introduction of the additional two significant parameters definitely makes the dynamic regression model better than a univariate time series model.

The dynamic regression approach provided better forecasts than the SSARR model but it did not take into account the snowmelt effect from the upper regions of the Humber River. The model used was linear in nature and nonlinear hydrologic effects, if any, could not be captured by the model.

2.3.3 Rainfall-Runoff Routing Model

The third model to be tried by the WRMD was an in-house routing model developed by its own engineers. The model is based on a series of water balance equations that were organized and put in three EXCEL spreadsheets to model three different parts of the basin.

The model used on the Upper Humber River at Black Brook mainly deals with effective rainfall since there is no upstream runoff or other explanatory variables to be considered. The effective rainfall is defined as the part that reaches the land surface. In this model, interception is the only loss considered. The amount of interception in this model is defined by a constant value. When the rainfall is greater or equal than the interception, the effective rainfall is equal to the observed rainfall minus the interception. Otherwise, the effective rainfall is equal to zero (Rollings 2008). This can be expressed by Eq. 2.7:

$$\begin{aligned}\text{Effective Rainfall} &= \text{Observed Rainfall} - \text{Interception} \quad \text{if Observed Rainfall} \geq \text{Interception} \\ &= 0 \quad \text{if Observed Rainfall} < \text{Interception)}\end{aligned} \qquad \text{(Eq. 2.7)}$$

After the calculation of effective rainfall, the daily net rainfall at Black Brook is computed by multiplying the effective rainfall by the drainage area and then converted to the units of cubic meters per day. This is shown in Eq. 2.8

$$\text{Net Rainfall} = \frac{\text{Effective Rainfall} \cdot \text{Drainage area}}{3.6 \times 24} \qquad \text{(Eq. 2.8)}$$

Where Net Rainfall is in unit of m^3/d

Effective Rainfall is in unit of mm/s

Drainage area is in units of km^2

The flow at Black Brook is then calculated in two ways. The first method is used when the observed flow at Black Brook of 1 day before is less than that of 2 days before. This is given by Eq. 2.9

$$\text{Flow}[t] = \frac{\text{Flow}[t-1]^2}{\text{Flow}[t-2]} + k1 \cdot \text{Net Rainfall}[t] + k2 \cdot \text{Net Rainfall}[t-1] + k3 \cdot \text{Net Rainfall}[t-2]$$

(Eq. 2.9)

Where [t-1], [t-2] is 1 day before and 2 days before, respectively;

k1 is runoff coefficient for rainfall of current day = 0.1;

k2 is runoff coefficient for rainfall of 1 day before = 0.6; and

k3 is runoff coefficient for rainfall of 2 day before = 0 .

The second method is used when observed flow at Black Brook of 1 day before is greater or equal than that of 2 days before. Then, the recession coefficient for high flows is introduced. This is shown in Eq. 2.10

$$\text{Flow}[t] = r \cdot \text{Flow}[t-1] + k1 \cdot \text{Net Rainfall}[t] + k2 \cdot \text{Net Rainfall}[t-1] + k3 \cdot \text{Net Rainfall}[t-2]$$

(Eq. 2.10)

Where r is recession coefficient for high flows (= 0.5)

The flow at Reidville is calculated simply by multiplying the flows at Black Brook by 2.5. Then the flows at Reidville, flows at Indian Brook, flows at Sheffield Brook, outflows of Hinds Lake,

and flows at Lewaseechjeech Brook are used to estimate the water level of Grand Lake, since all of them flow into the lake. In the second spreadsheet of the routing model, the net inflow of Grand Lake is calculated by the sum of the 5 sources multiplied with their coefficients minus the observed outflows. This is given by Eq. 2.11

Net Inflow of Grand Lake = 0.235 · Flow at Reidville + 2 · Flow at Indian Brook + 2 · Flow at Sheffield Brook + 1 · Outflow of Hinds Lake + 4 · Flow at Lewaseechjeech (Eq. 2.11)

The water level is also simply calculated as the observed water level of 1 day before adjusted by the net inflow. This is given by Eq. 2.12

$$\text{Water Level of Grand Lake}[t] = \text{Water Level of Grand Lake}[t-1] + \frac{\text{Net Inflow}[t-1]}{\text{Area of Grand Lake}} \quad \text{(Eq. 2.12)}$$

Where the area of Grand Lake is 467 km^2.

The third spreadsheet of the routing model deals with the water level of Deer Lake and its outflow which is also the streamflow at the Humber River at Village Bridge. In this model the inflow of Deer Lake is from four sources: flow from Reidville, local inflow below Grand Lake, local inflow to Deer Lake, and outflow of Grand Lake. Among them, flow at Reidville and outflow of Grand Lake are observed data. The local inflow below Grand Lake and local inflow to Deer Lake are both related to the flow at Reidville as defined by the routing model. The model calculates the water level of Deer Lake according to Eq. 2.13 and Eq. 2.14.

Net Inflow [t-1] = Flow at Reidville[t] + Local Inflow below Grand Lake[t]+ Local Inflow to Deer Lake[t] + Outflow of Grand Lake[t] (Eq. 2.13)

Where $$\text{Local Inflow below Grand Lake} = \frac{199}{2108} \cdot \text{Flow at Reidville}$$

$$\text{Local Inflow to Deer Lake} = \frac{640}{2108} \cdot \text{Flow at Reidville}$$

$$\text{Water Level of Deer Lake[t]} = \text{Water Level of Deer Lake [t-1]} + \frac{\text{Net Inflow [t - 1]}}{\text{Area of Lake}} \quad \text{(Eq. 2.14)}$$

The flow at Village Bridge (outflow of Deer Lake) of the current day is calculated based on the current water level of Deer Lake according to Eq. 2.15

$$\text{Flow at Village Bridge[t]} = 251.5 \cdot \text{Water Level of Deer Lake[t]} - 1092 \quad \text{(Eq. 2.15)}$$

In the three spreadsheets, mean absolute error is calculated by comparing the calculated flows and the observed flows. The model was found to only work well for the Lower Humber at Deer Lake and Village Bridge. For the upper part of the basin especially at Black Brook, the model performed quite poorly because snowmelt from the upper part of the basin was not taken into account.

In addition to the problems of accuracy of the forecasts with the routing model developed by the WRMD, there was also a lack of proper documentation of the calibration of the model. Many of the parameters used were subjectively obtained. Therefore the WRMD has currently abandoned this model.

Chapter 3

Artificial Neural Networks (ANN)

In the last chapter, the three models that have been used by the WRMD for flow forecasting along the Humber River were reviewed. This chapter will provide a review of the use of artificial neural networks in general and their use in river flow forecasting in particular.

3.1 Background and General Features of ANN

In the last 15 years, Artificial Neural Network (ANN) based model has been widely applied within the field of hydrological modelling (Li et al. 2008; Dawson et al. 2006; Campolo et al. 2003; Danh et al. 1999). An ANN is an advanced computation and simulation model which has been widely used in many areas of research and practical applications. An ANN operates like a human brain to provide a modelling route that can link the input X to the output Y. An ANN consists of neurons and connections similar to a biological neural system. In real life, the things people see, hear, and feel come into the brain and become the experiences in their memories. These experiences will tell them what to do better in the future when they are doing similar things. Like the human brain, the function of an ANN in engineering application is usually to learn the relationship between the inputs and outputs from a given set of data so that it can be used to predict future output values from new given input values (Kneale et al 2005).

An ANN has several advantages over traditional modelling techniques such as regression:

1. ANN has the ability to use field recorded data directly without simplification, unlike regression analysis, which requires an assumption of a functional form of the regression equation in advance.
2. ANN models can simultaneously determine the effects of fixed and random input variables on the response variable;
3. Trained ANN models are able to generate a predicted value for the response variable for any reasonable combination of input variables;
4. Valuable insight into interactions between variables, as well as the contribution of random variables to the response variable, can be gained (Baxter et al. 2004); and
5. An ANN can do parallel computations and can simulate a nonlinear system which is hard to describe by traditional modelling methods (Kerh and Lee 2006).

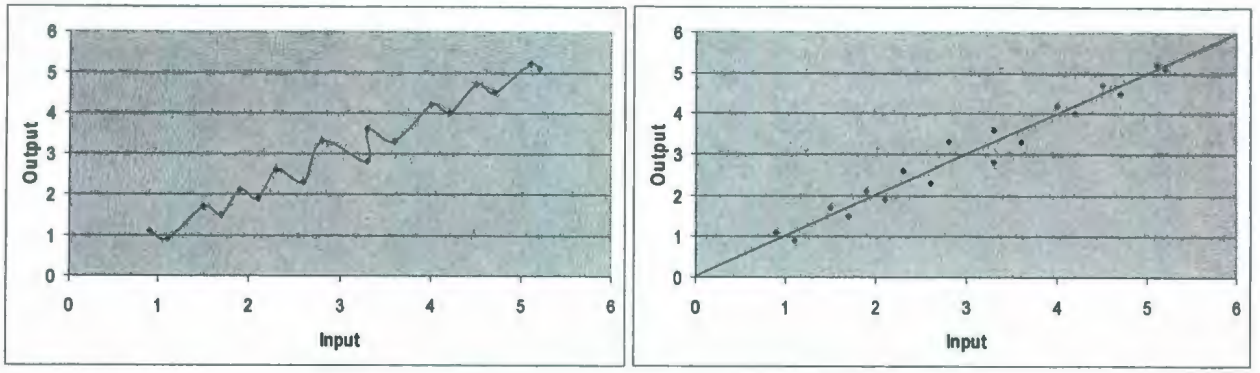
Using the modern computer technology, ANN can perform quick and efficient simulations of very complex problems and very large data sets. The solution from an ANN can be considered as a reference for further system modelling (Kneale et al 2005; Dawson et al 2006).

ANN is not without disadvantages. The two main disadvantages are computational time and the danger of "overfitting". Unlike regression analysis where the coefficients can be efficiently calculated by matrix algebra regardless of the number of data points or variables, an ANN usually requires a trial and error approach. One is never sure whether a unique optimal model has been obtained. The second disadvantage is that an ANN, like the human brain, has the defect of over-memorizing (also called "overfitting"). The ANN model can be over trained with the training data, and thus lose its power of generalizing to forecast any future data. This situation usually

happens when the training process is too long or too many hidden neurons are included in the model (Kneale et al 2005).

The number of processing units in the input and output layers is fixed according to the number of variables in the training data and is specific to each individual problem depending on the number of predictors. But the selection of an optimal number of hidden layers and hidden units will in all cases depend on the nature of the application. It is suggested by intuition that 'more is better' since a larger architecture will extend the power of the model to perform more complex modelling operations. But there is an associated trade-off between the amount of training involved and the level of generalization achieved. The use of large hidden layers can also be counter productive since an excessive number of free parameters encourages the overfitting of the network solution to the training data and so reduces the generalization capabilities of the final product.

For example, a group of data are known to follow the linear relationship of Y=X, but the training patterns fluctuate around the straight line as shown in Figure 3.1b. If there are too many hidden neurons in the network or the training takes a long time, the network may develop false surface features (like a pulse), as shown in Figure 3.1a. In this situation, the model not only fit the signal but also the noise from the training patterns. Although it can accurately describe the training patterns, it loses its ability of representing any further data. The objective of ANN is to generate a model that can fit generalized data rather than a certain group of data. Therefore, selecting an appropriate set of training patterns and configuration of the hidden neurons become critical.



(a) (b)

Figure 3.1. Incident of overtraining

The traditional way to determine an appropriate number of hidden neurons is by trial and error, although some software packages provide general guidelines to prevent overfitting.

3.2 Mechanism of ANN

Neurons and connections are the two basic components of an ANN architecture. The objective of an ANN is to figure out the neurons arrangement and connection weights. Neurons are usually arranged into three kinds of layer: input layer, hidden layer, and output layer. Figure 3.2 shows the architecture of a standard three layer neuron network. The neurons not only receive input signals but also output information with a particular strength to the input paths of other neurons through connection weights. All the neurons compute their outputs using their output functions and then the results may be put through their neighbouring neurons for the next step of processing. For each neuron, an intermediate value that comprises the weighted sum of all its inputs $I = \sum W_{ij}X_i$ is computed first (where X is the input value, W is the weight of each input value, I is the weighted summation, i is the number of the input source, and j is the number of the

target neuron). This value is then passed through a transfer function f(I), which performs a non-linear 'squashing operation' to calculate an activation level of this neuron. The microstructure of the neuron processing is shown in Figure 3.3. The transfer functions can be selected by the user. Most software packages have several common options such as logistic (sigmoid), linear, Gaussian, and hyperbolic tangent transfer functions available.

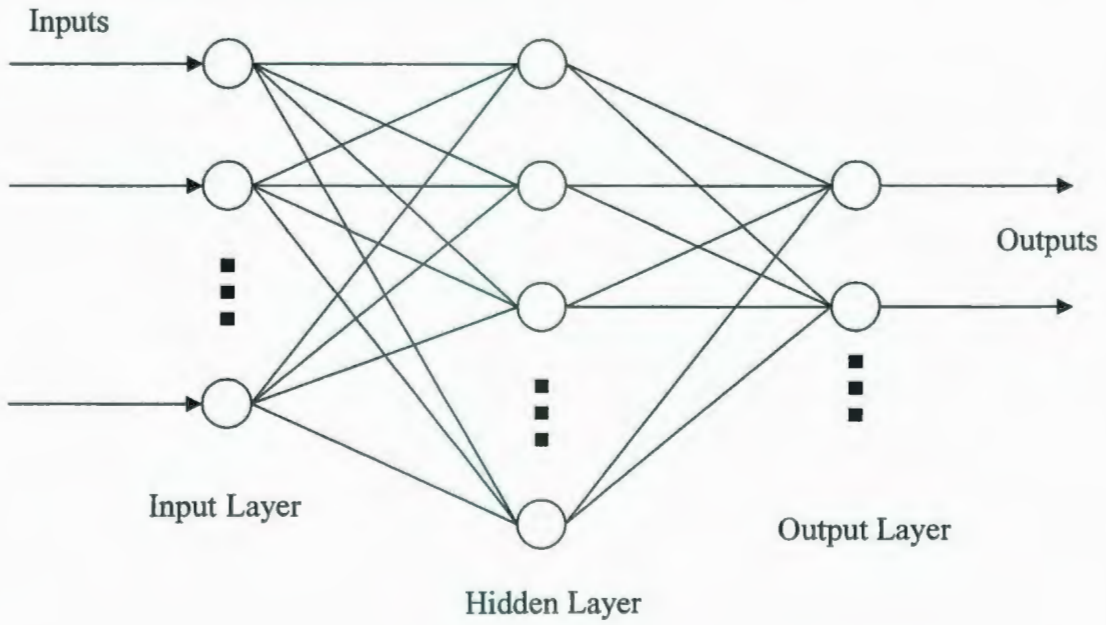


Figure 3.2 Architecture of a standard 3 layer neural network model

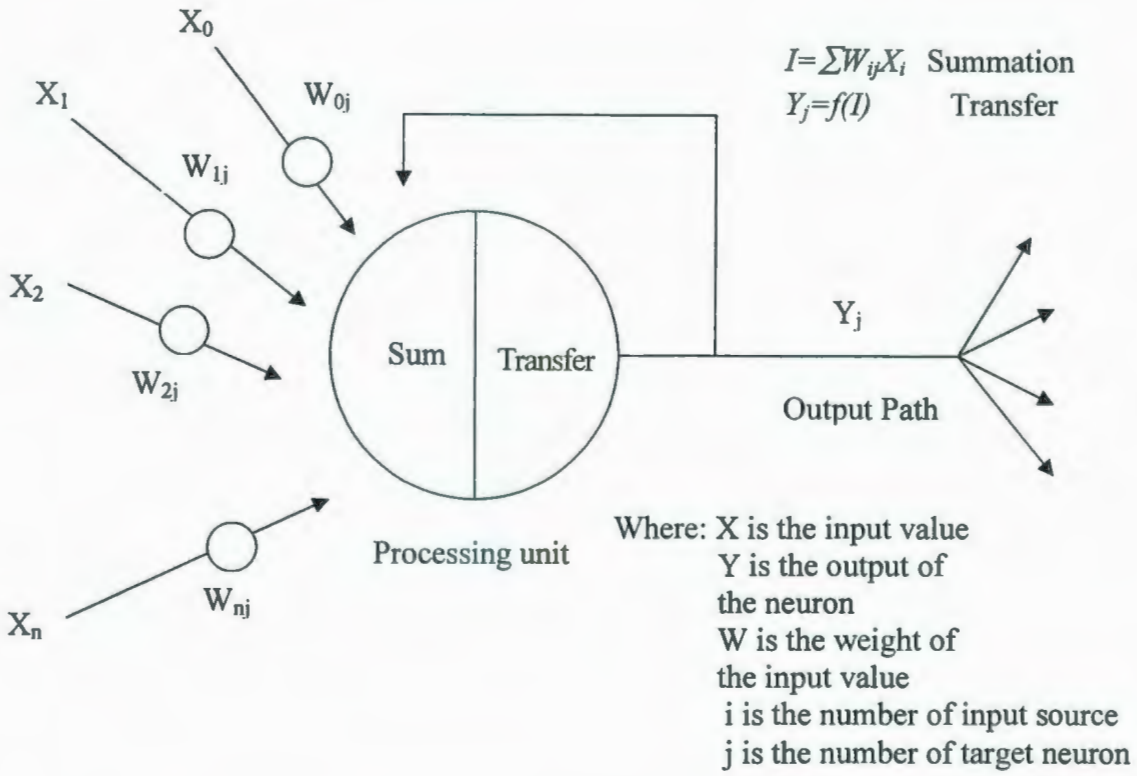


Figure 3.3 The microstructure of a neuron in the network

3.2.1 Transfer functions

Logistic (Sigmoid) – This function is found to be useful for most neural network applications. It maps values into the (0, 1) range. This function is always used when the outputs are categories.

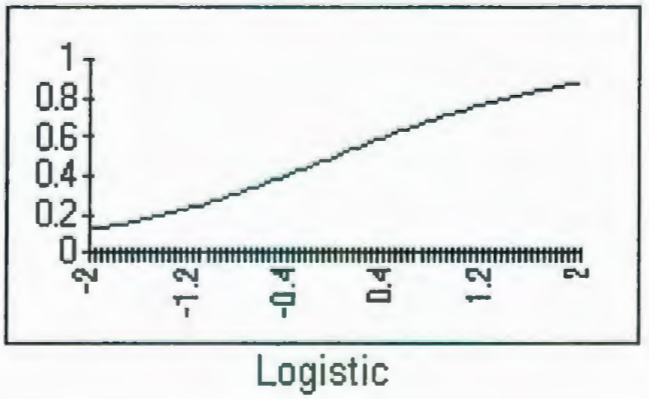


Figure 3.4 Plot of Logistic Function

Linear - Use of this function should generally be limited to the output layer. It is useful for problems where the output is a continuous variable, as opposed to several outputs which represent categories. Although the linear function lacks power for complex network modelling, it sometimes prevents the network from producing outputs with more error near the minimum or maximum of the output scale. In other words the results may be more consistent throughout the scale. If this function is used, it is better to use smaller learning rates, momentums, and initial weight sizes. Otherwise, the network may produce larger and larger errors and weights and hence will never reduce the error. The linear activation function is often ineffective for the same reason if there are a large number of connections coming to the output layer because the total weight sum generated will be high.

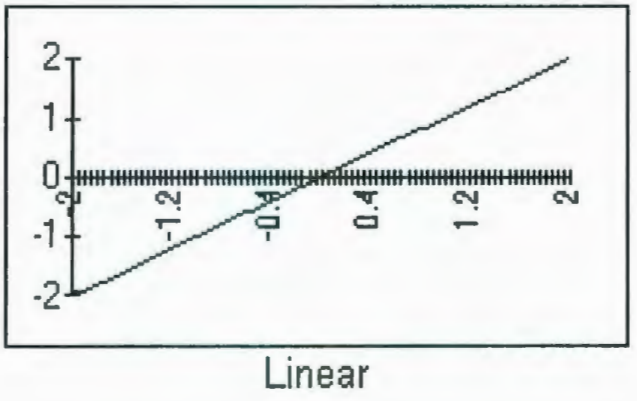


Figure 3.5 Plot of Linear Function

Tanh (hyperbolic tangent) – This function is not usually used during many projects. However, it is sometimes better for continuous valued outputs, especially if the linear function is used on the output layer. The scale of inputs of this function is [-1, 1]. Ward Systems Group (1993) has experienced good results when using the hyperbolic tangent in the hidden layer of a 3 layer network, and using the logistic or the linear function on the output layer.

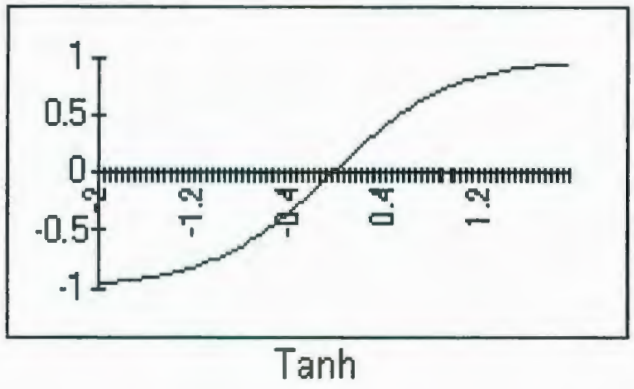


Figure 3.6 Plot of Tanh Function

Gaussian - This function is unique, because unlike the others, it is not an increasing function. It is the classic bell shaped curve, which maps high values into low ones, and maps mid-range values into high ones. There is not much about its use in the literature, but Ward Systems Group (2000) has found it very useful in a small set of problems. It is suspected that meaningful characteristics are not found at the extreme ends of the sum of weighted values. This function produces outputs in the range of [0,1].

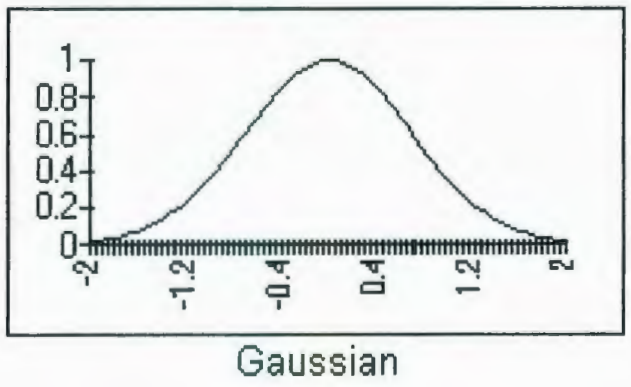


Figure 3.7 Plot of Gaussian Function

3.2.2 Network Training

To enable the ANN to represent a set of data, there is a need to adjust the connection weights or network structure. This adjusting procedure is called "training". There are two general methods of training, supervised and unsupervised (Daniel 1991; Flood and Kartam 1994; Abrahart 2005; Jain and Deo 2006). In the former case, solutions (output) are provided associated with the example problems (input). The number of hidden neurons, topology of connections (network structure), and the weights of connections are adjusted by small amounts by some rules to reduce the error between estimated output and targeted solutions. The process is repeated many times until the error meet a specified tolerance. In the latter case, the training set consists of inputs only. This situation may be because appropriate solutions are not available or because there is a desire to let the system identify the outputs by itself. The network can organize the inputs in any way it wishes. The processing elements can be organized in clusters with either competition or cooperation between the clusters occurring. Information usually reverberates around the network until some convergence criteria is met. The specifics of the algorithm for network training will be discussed after introducing some of the new terminologies used in the ANN training algorithm.

3.2.3 New Terminologies Used in ANN

ANN is a relatively new statistical tool in a sense that it has only been widely used in the last two decades or so. It has no predetermined functional relationship, and no exact rules for developing the ANN. Some new terminologies are therefore required:

- Training pattern: one set of inputs associated with one set of outputs. For example, in forecasting daily river flow, the physiographical and meteorological data and the associated river flow of one day is considered as one pattern.
- Epoch: one pass through the whole training patterns before one weight update is made.
- Learning rate (β): It is the rate of change of weights after one iteration. A high value is suggested at the start to speed up the progress. If it is too high, an oscillatory state may result.
- Momentum factor (α): It controls the speed of error correction and determines the effect of the previous weight change on the current change, which can take a solution trapped in the local minimum out of it.

Other terms will be described as needed.

3.3 Categories of ANN Model

Based on the various architectures and training algorithms, an ANN can be divided into different categories, such as BPNN (back propagation neural network), GRNN (general regression neural networks), and so on. In this thesis, only BPNN and GRNN will be considered as these two are applied for river flow forecasting (Campolo et al 2003; Kerh and Lee 2006; Li et al 2008).

3.3.1 Backpropagation Neural Network (BPNN)

"Backpropagation" is the most popular 'default' training algorithm for ANN, and it has been used by many researchers for daily flow forecasting. ANNs trained using backpropagation are also known as "feedforward multi-layered networks trained using the backpropagation algorithm" (Abrahart 2005). The mechanism and process of the standard 3-layer BPNN is

described as below. Firstly, the given data are stored in the input neurons. The input neurons then transmit these values across the links to the hidden neurons. On each link there is a weight used to multiply transmitted values. The weighted sum associated with the neuron bias is then put through a simple function (transfer function or activation function) to generate a level of activity for the hidden neuron. The activation levels of hidden neurons are then transmitted through their outgoing links to the neurons in the output layer. As before, the values are weighted and summed during transmission. Then, the summed value is put through an activation function to get an activation level of the output neurons, which is the final solution of the network. It provides an efficient computational procedure to evaluate the performance of the network.

After variables are loaded into a neural network, they must be scaled from their numeric range into the numeric range that the neural network can deal with efficiently. The common numeric ranges for the networks to operate in are from 0 to 1 denoted (0, 1) and minus one to one denoted (-1, 1),

The activation function usually used for back propagation is a sigmoid function (Kneale et al 2005) as described earlier.

$$f(I)=\frac{1}{1+e^{-I}} \qquad \text{(Eq. 3.1)}$$

$$\text{where } I_i = \sum_{j=1}^{n} w_{ij} x_j$$

The sigmoid function is found to be useful for most neural network applications. It maps values into the (0, 1) range.

The weight updates are based on a variation of the generalized delta rule (Kneale et al 2005).

$$\Delta w_{ij} = \beta E f(I) + \alpha \Delta w_{ij}^{previous} \qquad \text{(Eq. 3.2)}$$

where E is the error;

α is momentum factor;

β is learning rate.

In BPNN, errors of the current layer are calculated based on the errors of the former layer. For example, the error of output layer is $E_{ij}^{output} = y_j^{desired} - y_j^{actual}$, then the error of the hidden layer can be calculated according to

$$E_i^{hidden} = \frac{df\left(I_i^{hidden}\right)}{dI} \sum_{j=1}^{n} \left(w_{ij} E_j^{output}\right) \qquad \text{(Eq. 3.3)}$$

This is an operation that errors are propagated backwards across the network. Hence it is named 'backpropagation neural network'.

Other than the standard connections, there are some other kinds of backpropagation neural networks that can be built to solve different types of problems:

a) Jump Connections: This is the type of backpropagation network in which every layer is connected or linked to every previous layer. Three, four, or five layers of jump connection network can be selected. This network architecture may be useful when working with very complex patterns.
b) Recurrent Networks: This type of network is known for its ability to learn sequences, so it is suggested for time series data. The input, hidden, or output layer of are fed back into

the network for inclusion with the next pattern, which means the features detected in all previous patterns are fed into the network with each new pattern. This network usually takes longer to train.

c) Kohonen Self-organization network: This type of network consists of two layers with the first layer being for the input and the second layer for the processing input patterns. The Kohonen network is able to learn without being shown the correct outputs in the training patterns and the network models the probability distribution of the input vectors.

Generally, it has been shown in many research and practice that the three layers backpropagation neural network with standard connections is sufficient for the vast majority of problems. The architecture of the BPNN is the standard 3 layer neural network shown in Figure 3.2.

3.3.2 General Regression Neural Network (GRNN)

GRNN (general regression neural network) is a type of supervised network which is known for quick training on sparse data sets. GRNN works by comparing patterns based on their distance from each other and it is usually applied to continuous function approximation with multidimensional inputs. It is found that GRNN can produce better solutions than backpropagation in many types of problems (not all) (Ward System Group 2000).

GRNN is comprised of three layers. The first layer consists of neurons of input variables. The number of neurons in the first layer is equal to the number of input variables. The number of hidden neuron in the second layer is equal to the number of training patterns because the input pattern should be compared in N dimensional space to all of the patterns in the training set to

determine how far in distance it is from those patterns. The number of neurons in the third layer is equal to the number of outputs. The output that is predicted by the network is a proportional amount of all of the outputs in the training set (Gourrion 2000; Ward System Group 2000). The proportion is based upon the distance between new patterns and given patterns in the training set. According to the Gaussian Kernel regression estimator, this proportion can be defined as (Savelieva 2004):

$$W_i(x,y) = \frac{\exp\left(-\frac{d_i^2}{2\sigma^2}\right)}{\sum_{j=1}^{n}\exp\left(-\frac{d_j^2}{2\sigma^2}\right)} \qquad \text{Eq. 3.4}$$

where d_i is the computed distance, and σ is the spreading factor or smoothing factor of the transfer function.

According to the mechanism of GRNN, computing the distance d_i of new patterns from the patterns in the training set is a critical step in GRNN. Two methods of computing this distance are usually introduced in GRNN (Ward System Group 2000).

1. Vanilla or Euclidean distance is defined as root of the sum of squared difference in all dimensions between the pattern and the weight vector for that neuron. This method is mostly recommended since it is proved to work best for GRNN

   For example, we have 2 points P, Q in n dimensional space. $P = (p_1, p_2, \ldots p_n)$ and $Q = (q_1, q_2, \ldots q_n)$

   The Euclidean distance is (Savelieva 2004):

$$d = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \qquad \text{Eq. 3.5}$$

2. City Block distance metric is the sum of the absolute values of the differences in all dimensions between the pattern and the weight vector for that neuron. This method is computed faster than the Euclidean distance, but is usually not as accurate (Ward system Group 2000).

GRNN is essentially a non-parametric regression network (Savelieva 2004). There are no training parameters like learning rate, momentum, and calibration interval as in BPNN. There is only a smoothing factor which is applied after the network is trained. The smoothing factor will be automatically computed in Neuralshell2 (Ward System Group 2000) if a test set is extracted from the data set for calibration. The success of GRNN networks is dependent on the selection of the smoothing factor.

In Neuralshell2, a well-known ANN software developed by Ward Systems (2000), there are 2 options for the calibration of a GRNN network.

1. Iterative: this option is usually used when all of the input variables have the same contribution on predicting the output, for example, if the input variables are the same type (for example flows in m^3/s). The smoothing factor computed by an iterative method represents general impacts of inputs on the outputs.
2. Genetic adaptive: the one computed by this option is a combination of smoothing factors. Each input variable has its own smoothing factor which represents the contribution of this input variable on predicting the output. The larger the factor for a given input, the more important that input is to the model at least as far as the test set is concerned. So, this

option is usually used when the input variables are of different types and some may have more impact on predicting the output than others. The overall smoothing factor can also be modified by adjusting the individual smoothing factor of each variable. The training by the genetic adaptive methods takes longer than using the iterative method (Ward System Group 2000).

After the application of the training algorithm, it is necessary to determine how well the network performs on input patterns for which it was not trained. This process is basically a test to see how well the network has discovered the hidden features and sub-features in the training cases. The most usual way to deal with it in many neural network software is to divide the collected data into a training set, a test set, and a validation set. The training set is used for the network to learn the features of the data. Hence it is essential to have enough training cases to train the network. All the test cases that are collected in a test set will function as guidelines for the network. The training and testing process are carried out simultaneously to avoid over-fitting of the data. This is done by propagating the trained network on the test set and then the error of the test set is calculated. After the average error of the test set has stopped fluctuating for an optimum number of epochs, the learning process will be stopped. After learning has stopped, the "programmed" network will be applied on a validation set which has not been used in previous procedure. A well "programmed" network will usually give satisfactory results on this set of data. The extracting of these three sets from the known data is usually random to make sure all three sets have the features and sub-features of the data. In hydrology, the data are periodic. Therefore it is more appropriate to divide the hydrologic data according to years as this means that each set of data will contain the features of different seasons each year.

Chapter 4

Model Calibration

Most simulation models have several parameters that the user can adjust for different cases or purposes of use. The results produced by the models are usually different when using different values of parameters. In order to have the model represent as accurately as possible the system being modelled, there is a need to determine these model parameters by using known system inputs and responses. The process of determining the optimal value of these parameters is called calibration. For a rainfall-runoff model, inputs to the system would be variables such as rainfall amount for the day and for a previous day, temperature, and runoff from a previous day. The output would be the runoff for the day ahead. In order for the model to produce outputs that match the observation, the model must be properly calibrated.

4.1 Trial and Error and Automatic Calibration Methods

Traditionally most engineers conduct the calibration of hydrologic models by using a trial and error approach. This method is easy to understand but the results are not always satisfactory unless the modeller is very lucky and/or experienced. When doing trial and error calibration, the adjustment of the parameters is usually done one at a time. Parameters are optimized separately from each other. Each parameter is optimized by setting it to different levels within a defined range and then the goodness-of-fit of the output is checked. The goodness-of-fit is normally based on some numerical criteria and a graphical match. The Nash–Sutcliffe efficiency, average absolute difference |E|, and mean squared error, are some typical numerical criteria for goodness-of-fit testing. The adjustment of one parameter will stop when no improvement is made in the

goodness-of-fit. Each parameter follows the same process to find its optimal value without considering effects of other parameters. Although this method is simple and accepted by many engineers, it has some significant disadvantages which may cause unsatisfactory and non-optimal results.

1. Since each parameter is independently adjusted, this method is not able to consider the interactions between the parameters. A single parameter at its optimized level may improve the model but this may not be so when two or more parameters are applied together.
2. Parameters cannot be adjusted together simultaneously. They must be conducted one by one so that it requires a great deal of time. That is also the reason that the interaction between parameters cannot be determined.
3. This method cannot obtain the global optimal solution because the parameter- interaction problem.
4. It is difficult to exactly know when to terminate the calibration since it cannot be certain if the global optimal solution has been achieved.

The trial and error method is often used in simple models which have a straightforward structure and has very few parameters with the key assumption that the parameters do not interact with one another.

With the advancement of computer technology, newer advanced methods that use automatic calibration procedures based on computer programs, for example of Least Squares and Maximun Likelihood (Sulistiyono 1999) have been developed which speed up the process but very few of

them are capable of dealing with the parameter- interaction problem. Another disadvantage of these automatic calibration methods is that they do not help the user understand the behaviour and contribution of each model parameter and their possible interactions.

4.2 Design of Experiments Methodology

Design of Experiments (DOE) methodology is widely used as a preliminary step in many research and industrial processes (Myers and Montgomery 1995). It is a systematic process to observe and identify the relationship between the changing of input variables and the resulting change of output responses. Through the DOE process, it is possible to determine one or a group of appropriate input combinations that produces outputs that achieve a particular goal. This characteristic of DOE provides another method of model calibration that is more informative than either the trial-and-error or automatic methods. The DOE approach is well documented and simple, and many researchers and modellers have started to use DOE as a tool to optimize their process or calibrate their models. Besides this, DOE also provides the benefit of conducting a proper experimental design and learning about the behaviour of the model and parameters. Some of the key benefits of DOE include:

1. Capability of dealing with parameter interactions: not only effects of single model parameters but interactions between two or more model parameters can be statistically tested objectively.
2. It is efficient because DOE can help locate the best range of inputs for further analysis with very few trials.

4.2.1 Factorial Design

For a factorial design, every level of every variable is paired with every level of every other variable as a combination, and each combination is considered as an experiment (Johnson and Leone 1997). In model calibration, the combinations of factor levels are the input variables of the model. Then the response obtained as a result of each combination is the output of the model. Factorial design is a very general kind of design. It can handle any number of factors or model parameters with any number of levels. The most efficient designs however are the 2-level factorial and 2-level fractional factorial designs. The fractional factorial design is usually used to screen the factors (model parameters) when there are many factors are involved. The total number of experiments to complete a design is based on the number of factors and the number of levels of each factor. For example, the number of required experiments of a 2-level full factorial design is equal to two to the power of the number of factors (2^k, k=the number of factors). For a fractional factorial design, the number of runs can be determined as 2^{k-p}, where p is the fraction of the number of runs. For example, if k=6, and p=1, then only 32 runs are required instead of 64. The choice of the fraction must however depend on the resolution required. Further details on full factorial and fractional factorial designs are available in standard text books such as Myers and Montgomery (1995). The result of the experimental design consists of three parts: parameter effect estimation, model fitting, and optimization. Only when the first two results are satisfied, can optimization be carried out.

4.2.2 Response Surface Methodology (RSM)

Response surface methodology (RSM) is a well known method for optimizing processes based on a polynomial surface analysis. This method is widely used in quality improvement, product

design, uncertainty analysis, and so on. The mechanism of RSM is to use mathematical and statistical techniques to generate a polynomial function of several variables ($\eta = f(x_1, x_2, \ldots, x_k)$). This function is then called a response surface (Myers and Montgomery 1995). The objective is to optimize the response as required, for example, maximize, minimize, or get to a target value. Using RSM for model calibration can not only help us determine the optimum combinations of factors and their levels that will satisfy a set of desired specifications but also describes how a specific response is affected by changes in the level of the factors over the specified levels of interest. This function will allow a modeller to get a better understanding of the contribution that each factor makes.

Usually the form of the relationship between the response and independent variables is unknown. So, the first step of RSM is to find a suitable approximation for the true relationship between response and variables (Myers and Montgomery 1995).

If the response is well modelled by a linear function of the independent variables, the response looks flat if it can be plotted. Then the approximating function is the first-order model (linear):

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \varepsilon \quad \text{(Eq. 4.1)}$$

When there is significant curvature in the response surface then a nonlinear model is warranted. A polynomial of higher degree must be used, such as the second-order model:

$$Y = \beta_0 + \sum \beta_i \cdot x_i + \sum \beta_{ii} \cdot x_i^2 + \sum \beta_{ij} \cdot x_i \cdot x_j + \varepsilon \quad \text{(Eq. 4.2)}$$

Where ε is the error between the approximation surface and the actual response.

Although models with higher degrees can be applied for the approximation, the second-order model is nearly always adequate if the surface is "smooth". (Myers and Montgomery 1995). The goodness-of-fit of the RSM is also examined by ANOVA. If the fitted surface is an adequate approximation of the true response function, the R-squared value of the approximated response should be close to 1. Then analysis of the fitted surface will be approximately equivalent to the analysis of the actual system (within bounds).

4.2.3 Central Composite Design (CCD)

Central Composite Design (CCD) is formed from the two level factorial designs with additional points that allow the coefficients of a second-order model to be estimated (Myers and Montgomery 1995). In Central Composite Design, each factor varies over five levels. Besides the two levels in factorial bases, at least one central point and two axial points are introduced into the design. A central point is at the middle between two levels of each factor in factorial design. Axial points are points on the coordinate axes at distances "α" from the design center. The value of "α" is usually selected to make the CCD rotatable. It is calculated as the fourth root of 2 to power k ($\alpha = \sqrt[4]{2^k}$, where k is the number of factors). For an example, for a central composite design with 3 factors x_1, x_2, and x_3, the experiments are done on the following points. The structure of these points in 3D form is also shown in Table 4.1 below.

Table 4.1: Example of a 3-factor rotatable CCD

runs	x_1	x_2	x_3
1	-1	-1	-1
2	-1	-1	+1
3	-1	+1	-1
4	-1	+1	+1
1	+1	-1	-1
6	+1	-1	+1
7	+1	+1	-1
8	+1	+1	+1
9	- α	0	0
10	+ α	0	0
11	0	+ α	0
12	0	- α	0
13	0	0	+ α
14	0	0	- α
15	0	0	0

$\alpha = \sqrt[4]{2^3} = 1.682$

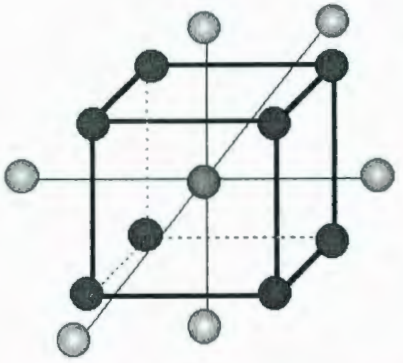


Figure 4.1: 15 runs of 3-factor rotatable CCD showed in 3D plot

CCD has the advantage that it can be done on the base of a 2 level factorial or fractional factorial design in stages. The factorial part can be a fractional factorial as long as it is of Resolution V or greater so that the 2 factor interaction terms are not aliased with other 2 factor interaction terms. This advantage makes the CCD more efficient than other RSM designs when many factors are

introduced. Also, rotatability is the property relating to the precision of the predicted response value. An experimental design is said to be rotatable if the variance of the estimated response depends on the distance from the design center rather than the direction (Cornell 1990; Myers and Montgomery 1995). In other words, rotatability ensures that the error in prediction stays constant around the design. This property is not achieved by many other response surface methodologies.

4.2.4 Steps in Using DOE for Model Calibration

When applying DOE in general, or for model calibration in particular, the procedures used are as follows:

1. Determine the parameters to be used in the model and their ranges. Before calibrating a model, it is necessary to know what parameters the model uses and the lower and upper limits of each parameter that can be independently adjusted. In some cases, the input parameters can be either numerical or categorical. The ranges of these two kinds of parameters are then defined differently. For example, the range of temperature can be defined as from -20°C to 30°C. The range of "if temperature is above zero" could be "yes" or "no".
2. Determine the objective functions. The objective functions, which are the outputs for DOE analysis, are used to evaluate the results of the model. Usually they are goodness-of-fit measures. Generally, the more objective functions used, the more precise the optimization. The objective functions used in this research will be described later.
3. Choose the experimental design. In DOE, many methods are available. These include the 2-level factorial and fractional factorial designs, Central Composite design (CCD), Box-Behnken design (BBD), and so on. The selection of the design mostly depends on the

model characterization and complexity. For example, if the model has many parameters, a fractional factorial design is suggested for screening of important parameters. If the model is highly nonlinear, or there is a significant curvature in the response surface, CCD and BBD or others may be applied.

4. Estimate the effects of parameters and parameter-interactions. The effects of both parameters and parameter-interactions can be estimated by using standard regression analysis. The effects show the different contributions of each parameter and the interaction effects show the effect of one parameter as another parameter is changed. The significance of each effect can be determined by using analysis of variance (ANOVA) or by other means such as the normal probability plot or Pareto plot. This will help the researcher decide which parameter is more important and also determine which parameters interact and have to be jointly considered.

4.3 Calibration of ANN Models by DOE

When the ANN is trained by backpropagation, values of several internal ANN parameters have to be determined. They are the learning rate β, momentum α, number of hidden neurons, and calibration interval. Selection of the values of these parameters often affects the performance of the model. As mentioned earlier, these parameters are usually set based on experience or are changed one parameter at a time to see if there is improvement of the model. In this thesis, to investigate the most appropriate combination of parameters, a design of experiment (DOE) methodology is applied. Hence, 4 numerical factors need to be calibrated for the ANN model. Anticipating the possible nonlinear relationship among the factors (model parameters), the CCD (central composite design) is selected. In order to calibrate the model, data at Black Brook are

taken as an example. All the possible climatic variables are considered to be included in this part. The further analysis of variable selections will be discussed in Chapter 5. Then, the inputs for the model at Black Brook are:

- QB_{t-1}: Daily flow at Black Brook of 1day ahead;
- QB_t: Daily flow at Black Brook of current day;
- TB_{t-1}: Air temperature at Black Brook of 1day before;
- TB_t: Air temperature at Black Brook of current day;
- PB_{t-1}: Total precipitation at Black Brook of 1day before;
- PB_t: Total precipitation at Black Brook of current day;
- DD_{t-1}: Cumulative degree days up to 1day before; and
- DD_t: Cumulative degree days up to the current day.

4.3.1 Parameter (factor) ranges

Since the momentum is theoretically in the range [0, 1], 0 and 0.99 are used as the lower and higher limits of momentum. For the learning rate the lower and upper limit used are 0.01 and 1, respectively. For the number of hidden neurons, the lower and upper limits used are 20 and 60. The reason for choosing this range is because the suggested number of hidden neurons by Neuralshell2 is 43, which is almost at the center of the range from 20 to 60. In addition, the range in CCD is usually entered in terms of α. The value is 2 for 4 numerical factors which make the 5 levels of this parameter to be: 20, 30, 40, 50, and 60. This distribution meets the integer requirement of the numbers of hidden neurons. For the calibration interval, it is suggested better to start with 200 (Ward System Group 2000). Therefore a lower limit of 20 and an upper limit of 200 are used for this parameter. The 5 levels, 20, 65, 110, 155, 200, also meet the integer requirement of this parameter. (Table 4.2).

Table 4.2 Levels of each factors selected for CCD design

Factors	Level 1	Level 2	Level 3	Level 4	Level 5
A: Learning rate	0.01	0.26	0.51	0.75	1.00
B: Momentum	0	0.25	0.49	0.74	0.99
C: No. of hidden neurons	20	30	40	50	60
D: Calibration interval	20	65	110	155	200

4.3.2 Outputs or Responses

Neuralshell2 provides many indicators to check the accuracy of the fitted model; some of these indicators or goodness-of-fit measures are then used as the responses of DOE experiment for each combination of the input factors. The measures used are: Nash-Sutcliffe efficiency, correlation coefficient r, mean squared error, mean absolute error, and the percentage of outliers.

1. The Nash-Sutcliffe model efficiency coefficient.

The Nash–Sutcliffe model efficiency coefficient is often used to assess the predictive power of hydrological models. It can also be used to quantitatively describe the accuracy of model outputs beside hydrological discharges as long as there is observed data to compare the model results to. In other applications, the measure may be known as the coefficient of determination, or R^2. Nash–Sutcliffe efficiencies can range from $-\infty$ to 1. An efficiency of 1 ($E = 1$) corresponds to a perfect match of modeled discharge to the observed data. An efficiency of 0 ($E = 0$) indicates that the model predictions are as accurate as the mean of the observed data, whereas an efficiency less

than zero ($E < 0$) occurs when the model predictions are worse than what could be predicted by just using the mean of the sample case outputs. Essentially, the closer the model efficiency is to 1 the more accurate the model is. Nash–Sutcliffe efficiency is defined as:

$$E = 1 - \frac{\sum_{t=1}^{T} \left(Q_0^t - Q_m^t\right)^2}{\sum_{t=1}^{T} \left(Q_0^t - \overline{Q}_0\right)^2} \qquad \text{(Eq. 4.3)}$$

Where Q_0 is observed discharge, Q_m is modeled discharge, and Q_0^t is observed discharge at time t.

2. The correlation coefficient r (Pearson's Linear Correlation Coefficient)

This is a statistical measure of the strength of the relationship between the actual versus predicted outputs. The correlation coefficient can range from -1 to +1. The closer r is to 1, the stronger the positive linear relationship, and the closer r is to -1, the stronger the negative linear relationship. When r is near 0, there is no linear relationship. Pearson's correlation coefficient is written:

$$r_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \qquad \text{(Eq. 4.4)}$$

Where $\bar{x}$ and $\bar{y}$ are the means, s_x and s_y are the standard deviation

3. Mean squared error

This is the mean of the square of the actual value minus the predicted value. It is given by:

Mean squared error=$\sum$(actual - predicted)^2/N. (Eq. 4.5)

Where: N = total number of patterns or sample size.

4. Mean absolute error

This is the mean over all patterns of the absolute value of the actual minus predicted. It is given by:

$$\text{Mean absolute error} = \sum | \text{actual} - \text{predicted} | / N \qquad \text{(Eq. 4.6)}$$

Where: N has been previously defined.

5. Percentage of Outliers

The percentage of predicted values over 30% of the actual values is defined as the percentage of outliers. This could be an over predicted or an under prediction. The outlier is defined as:

$$\text{Outlier} = 100 \times |(\text{Predicted} - \text{Actual})|/\text{Actual} > 30\% \qquad \text{(Eq. 4.7)}$$

From the CCD, 25 combinations are required. Each combination is then used as input factors to the ANN model in Neuralshell2. The goodness-of-fit measures from the resulting ANN model for both training set and validation set are then used as the responses in the CCD. Table 4.3 shows the ANOVA table for the most significant factors in the model. It can be seen that all the factors are not statistically significant at the 5% level which means the model performance is not highly affected by the setting of parameters.

Table 4.3 ANOVA table for the significant factors for response of Nash-Sutcliff efficiency value

Response: Nash-Sutcliffe efficiency

ANOVA for Response Surface Reduced Quadratic Model

Analysis of variance table [Partial sum of squares]

Source	Sum of Squares	DF	Mean Square	F Value	Prob> F	
Model	0.000316	14	2.26E-05	1.113868	0.4416	insignificant
A	5.55E-05	1	5.55E-05	2.739733	0.1289	
B	8.28E-06	1	8.28E-06	0.408847	0.5369	
C	1.35E-06	1	1.35E-06	0.066815	0.8013	
D	1.82E-05	1	1.82E-05	0.898287	0.3656	
A^2	1.44E-05	1	1.44E-05	0.712038	0.4185	
B^2	4.88E-06	1	4.88E-06	0.240825	0.6342	
C^2	3.06E-05	1	3.06E-05	1.508023	0.2476	
D^2	1.03E-08	1	1.03E-08	0.000509	0.9824	
AB	4.94E-05	1	4.94E-05	2.435715	0.1497	

Table 4.4 10 best solutions estimated by DOE methodology for BPNN algorithm

	Parameters				Solutions by DOE					
Number	Learning Rate b	Momentum a	Hidden Neuron	Calibration Interval	If using degree days	Nash-Sutcliffe efficiency	Mean squared error	Max absolute error	percent over 30%	Desirability
1	0.53	0.37	35.00	132.50	0.892563	0.893252	126.486	137.645	0.945175	21.6327
2	0.54	0.37	35.00	132.50	0.892712	0.89341	126.311	137.758	0.945258	21.6656
3	0.54	0.44	35.00	132.50	0.891274	0.892138	128.016	137.061	0.944579	22.0801
4	0.56	0.61	45.00	121.32	0.891775	0.892432	127.43	138.381	0.944705	21.8838
5	0.56	0.61	45.00	121.12	0.891779	0.89243	127.426	138.388	0.944704	21.8679
6	0.56	0.61	45.00	123.74	0.891767	0.892483	127.439	138.388	0.944729	21.9001
7	0.55	0.62	44.96	116.61	0.891947	0.892523	127.227	138.473	0.944758	21.8934
8	0.57	0.60	45.00	125.32	0.891631	0.892363	127.603	138.475	0.944662	21.5261
9	0.54	0.62	45.00	106.14	0.892025	0.892531	127.134	138.456	0.944772	22.2473
10	0.50	0.57	35.00	132.49	0.889892	0.890941	129.644	137.073	0.943909	22.7747

For an optimal ANN model, all the goodness-of-fit indicators should be at the most desirable values so that an optimal combination of parameter setting can be obtained. This means that the Nash-Sutcliffe efficiency value should be maximized, the mean squared error should be minimized, the mean absolute error and percentage of outlier (over 30%) should also be minimized. The combinations of inputs that meet these optimal criteria are then found using the optimization routine in Design-Expert 7.1. Ten solutions were found that met the optimal criteria (Table 4.4). For the further analysis, the parameter combinations of the 10 solutions are then substituted back into Neuralshell2 to see if the indicators computed by the ANN match those obtained by DOE. According to the optimization results from DOE, the 10 solutions shown are very close. The performance indicators differed by less than 0.5% when different combinations of BPNN parameters are applied. It also proves that the model performance does not depend on the parameter settings.

In NeuroShell2, a set of default parameter values of BPNN will be given as long as the model inputs and outputs are decided. In this model, for a simple 3 layer backpropagation neural network, the default parameter values suggested by NeuroShell2 are: learning rate is equal to 0.05, momentum is equal to 0.5, Number of hidden Neurons is 43 and calibration interval is 110. When the default values of learning rate and momentum are used in the ANN model, the results are also close to those 10 solutions obtained by using optimization routine. The values of the BPNN parameters are hence set at the default values, because using the default settings will not affect the accuracy of the model but can facilitate the BPNN processing speed.

Same as the model at Black Brook, the performances of BPNN models at Reidville and Village Bridge differed slightly (the differences of Nash-Sutcliffe are also less than 0.5%) with or without calibration. In addition, the results of the models at Reidville and Village Bridge are already very good, hence there is not much room for improvement. Therefore, using the default settings suggested by NeuroShell2 is sufficient for the BPNN models.

Although the use of DOE methodology did not provide a significant improvement in the model compared to using the software's default values, the exercise did however provided insights into the importance of the various network parameters. Now it can be stated for certain that the 4 internal parameters of BPNN do not affect the response

Chapter 5

ANN Models and Results

In the last chapter, design of experiment (DOE) methodology was used as an objective tool to calibrate the parameters of the Artificial Neural Network (ANN) models. The objective of this chapter is to determine how well ANN models perform over traditional models in solving the streamflow forecasting problem and to compare the difference between the two types of ANN: backpropagation neural network (BPNN) and general regression neural network (GRNN). In addition, DOE methodology is used to determine the statistical significance of the input variables to the ANN models.

The following problems are considered:

1) Forecast the 1-day ahead streamflows of Upper Humber River at Black Brook station;
2) Forecast the 1-day ahead streamflows of Humber River at Reidville station; and
3) Forecast the 1-day ahead streamflows of Humber River at Humber Village station.

Neuralshell2, release 4.0 by Ward System Group, Inc. was used for all ANN processing. This software provides the BPNN algorithm with different training architectures as well as the GRNN algorithm. Also, this software can extract subsets of data from the original data for both training and testing. The training and testing are operated simultaneously in the learning process. Using this approach can help avoid over memorization of the data. The data provided by the Water Resources Management Division of the Newfoundland and Labrador Department of Environment and Conservation are from January of 1997 to June of 2008 with the exception of temperature data of 2001 and 2002 at Humber River near Black Brook as previously described.

Since the streamflow of the Upper Humber River at Black Brook can be influenced by snowmelt, snowmelt is thus an important factor that affects the magnitude of the flows at this station. In the literature, when dealing with snow covered areas or snowmelt problems, the Cumulative Degree Days index has been the most cited as a method to represent the degree of snowmelt from a snow covered region (e.g. Peeters 1998; Suzuki et al. 2003; Fleming and Ouilty 2007). In this thesis, only temperatures above or equal to 0°C have been taken into account in the calculation of the cumulative sums. Negative temperatures have been considered as equal to 0°C. Traditionally, the starting date for the cumulative sum is the 1st of January. But snow at the Upper Humber above Black Brook usually does not start to melt until April and temperatures before April are almost all below 0°C. The starting date of cumulating is taken to be when there are 5 consecutive days of temperatures above 0°C. Similarly, the end of the cumulating period is when there are 5 consecutive days temperatures are all below 0°C. Based on the 10 year database, the starting and end time of each year are calculated and then the average values selected. The average starting day is the 107th day (April 16th or 17th) of each year and the average ending day is the 310th day (November 5th or 6th) of each year.

Cumulative Degree Day of current day, DD_t= the sum of all the temperatures above 0°C from the starting date to the current day (Eq. 5.1)

As an example, consider the temperature data of 10 days. Assuming that the degree day starts from the first day, then the values of cumulative degree days are calculated as below:

Days	1	2	3	4	5	6	7	8	9	10
Temp	0	2	5	3	-1	0	5	7	3	5
DD_t	0	2	7	10	10	10	15	22	25	30

The cumulative degree day, DD_t, is a proxy for the amount of heating available to melt the snow pack. It is assumed that the larger the DD_t, the more snow will be melted. In the following sections, the ANN models for each of the stations will be developed in detail.

5.1 Modeling of the Humber River Flow at Black Brook (Upper Humber)

From the data availability at Black Brook, 8 factors are suggested as possible inputs for the ANN model and they are:

- QB_{t-1}: Daily flow at Black Brook of 1day before;
- QB_t: Daily flow at Black Brook of current day;
- TB_{t-1}: Air temperature at Black Brook of 1day before;
- TB_t: Air temperature at Black Brook of current day;
- PB_{t-1}: Total precipitation at Black Brook of 1day before;
- PB_t: Total precipitation at Black Brook of current day;
- DD_{t-1}: Cumulative degree days up to 1day before;
- and DD_t: Cumulative degree days up to the current day.

The output is the flow of the next day Q_{t+1}. This provides a 1-day ahead forecast. Since there are numerous missing data from the 2001-2002 data set, the whole data set was separated into two parts: 1997-2000 and 2003-2007. Data from 1997 to 2000 were used for training, and data from 2003-2006 were used for testing, and data of 2007 were used to verify the model. More data are

required for training and testing set to ensure that sufficient data are available to 'program' the network.

The selection and contribution of variables in a model is also of interest so that only variables or factors that significantly contribute to the goodness of fit of the model are used. In NeuralShell2, these contributions are shown depending on the type of ANN used. In GRNN, when the model is under training, a particular smoothing factor is distributed to each input parameter. After training begins the individual smoothing factors for each of the input variables are displayed. The input smoothing factor is an adjustment used to modify the overall smoothing to provide a new value for each input. At the end of training, the individual smoothing factors may be used as a sensitivity analysis tool: the larger the value for a given input, the more important that input is to the model, at least as far as the test set is concerned. Individual smoothing factors are unique to each network. The values are relative to each other within a given network and they cannot be used to compare inputs from different nets. In BPNN, the 'contribution factors detail' module can be used to provide a rough measure of the importance of each variable in predicting the network's output. These values are also relative to each other within a same network. The contribution factor is developed from an analysis of the weights of the trained neural network. The higher the value, the more the variable is contributing to the prediction or classification. However, these smoothing factors or contribution factors do not give an indication of statistical significance. For Humber River at Black Brook, the smoothing factors of each variable after the training of the GRNN are shown in Table 5.1. The 'contribution factors' when the ANN is trained by BPNN is shown in Table 5.2. Also shown are the rankings of importance in parenthesis. The rank of GRNN and BPNN are different because they have different training algorithms.

Table 5.1 Individual smoothing factors of GRNN at Black Brook

Input Variables	Individual smoothing factor	Rank
QB_{t-1}	0.61176	6
QB_t	2.97647	1
TB_{t-1}	1.30588	3
TB_t	2.45882	2
PB_{t-1}	0.56471	7
PB_t	1.00000	4
DD_{t-1}	0.17647	8
DD_t	0.76471	5

Table 5.2 Input strength of variables of BPNN at Black Brook

Input Variables	Input strength	Rank
QB_{t-1}	0.12291	4
QB_t	0.28491	1
TB_{t-1}	0.05107	8
TB_t	0.15993	2
PB_{t-1}	0.13329	3
PB_t	0.07294	7
DD_{t-1}	0.08971	5
DD_t	0.08523	6

From Tables 5.1 and 5.2, it is clear that in both ANN models the flow and temperature of the current day are the most important input variables to forecast the 1-day ahead flows. There is little agreement between the two models beyond the top two input variables. It does make sense that the current day flow would be an important variable as the flows from day to day are highly autocorrelated. Temperature as an input variable also makes sense as temperature affects snow melt.

The input strength and individual smoothing factors as stated earlier are not statistically based and another approach must be used to test these input variables or factors for statistical significance. In this regard, DOE methodology is used again.

GRNN Model

Firstly, the 3 least important input factors A: QB_{t-1}, B: PB_{t-1}, and C: DD_{t-1} are selected for the testing using a 2 level 3-factor factorial design. From the ANOVA, only factor A: QB_{t-1} is statistically significant at the 5% level (P-value < 0.05) using the Nash-Sutcliffe efficiency as the response. The ANOVA results for the Nash-Sutcliffe coefficient are shown in Table 5.3 and Figure 5.1. This indicates that the other 2 input factors are not statistically significant to the model. Next, the 4 more important factors are selected for testing. They are defined as A: TB_{t-1}, B: TB_t, C: PB_t, and D: DD_t. The factor QB_t has the highest individual smoothing factor value which means it is the most important factor in this model. So QB_t does not need to be tested, it should definitely be in the model. The results of the second step show that all of the 4 factors are statistically significant (Table 5.4). They should all be included in the model. After the 2-step DOE tests, it can be concluded that only factor PB_{t-1}, and DD_{t-1} are not statistically significant

out of the 8 possible input factors. The model input factors can be reduced to 6. They are: QB_{t-1}, TB_{t-1}, TB_t, PB_t, DD_t, and QB_t.

BPNN Model

The same DOE approach can be applied on the BPNN model as well. The 4 least important input factors A: TB_{t-1}, B: PB_t, C: DD_{t-1}, and D: DD_t are selected for the first test. The ANOVA results show that all the factors except factor A: TB_{t-1}, are statistically significant (Table 5.5). The second step test is then applied and the result shows that all 4 factors are all statistically significant (Table 5.6). Also, the dominant factor QB_t should definitely be included. Therefore, 7 factors out of 8 are statistically significant. The model input factors are then: QB_{t-1}, PB_{t-1}, TB_t, PB_t, DD_{t-1}, DD_t, and QB_t. According to the results of the 2-step tests of both GRNN and BPNN, the factors in the second step tests are all statistically significant if there is at least one factor in the first step that is statistically significant. Therefore, if it is found that some factors in first step test are statistically significant, it is not necessary to perform the second step test as the rest of the factors which are not tested in the first step will all be statistically significant.

Table 5.3 ANOVA of the 3 less important factors of GRNN at Black Brook

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F	Judgement
Model	0.000114	3	3.79E-05	10.42059	0.0232	significant
A	8.52E-05	1	8.52E-05	23.40124	0.0084	significant
B	2.08E-05	1	2.08E-05	5.716592	0.0751	n.s.
C	7.8E-06	1	7.8E-06	2.143937	0.2170	n.s.
Residual	1.46E-05	4	3.64E-06			
Cor Total	0.000128	7				

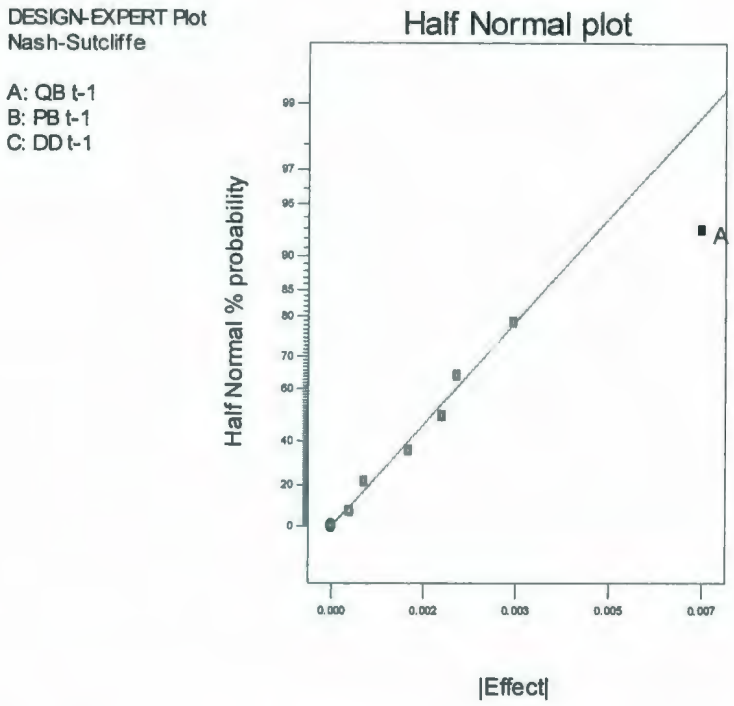


Figure 5.1 DOE effects plot for the 3 less important factors of GRNN at Black Brook

Table 5.4 ANOVA of the 4 more important factors of GRNN at Black Brook

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F	Judgement
Model	0.007133	8	0.000892	329.4097	< 0.0001	significant
A	0.000743	1	0.000743	274.3337	< 0.0001	significant
B	0.002965	1	0.002965	1095.322	< 0.0001	significant
C	0.001146	1	0.001146	423.3148	< 0.0001	significant
D	0.001897	1	0.001897	700.6844	< 0.0001	significant
AD	4.62E-05	1	4.62E-05	17.08299	0.0044	significant
BD	9.02E-05	1	9.02E-05	33.34213	0.0007	significant
CD	0.00021	1	0.00021	77.67516	< 0.0001	significant
ABD	3.66E-05	1	3.66E-05	13.5225	0.0079	significant
Residual	1.89E-05	7	2.71E-06			
Cor Total	0.007152	15				

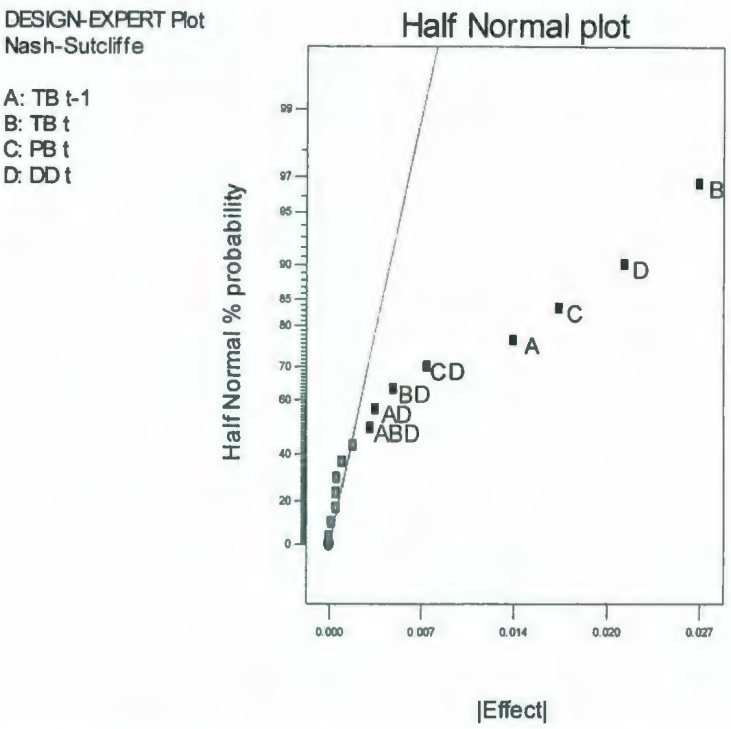


Figure 5.2 Effects plot for the 4 more important factors of GRNN at Black Brook

Table 5.5 ANOVA of the 4 less important factors of BPNN at Black Brook

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F	Judgement
Model	0.000175	7	2.5E-05	16.51779	0.0004	significant
A	6.25E-10	1	6.25E-10	0.000413	0.9843	n.s
B	7.61E-05	1	7.61E-05	50.3102	0.0001	significant
C	1.91E-05	1	1.91E-05	12.64973	0.0074	significant
D	1.91E-05	1	1.91E-05	12.64973	0.0074	significant
BC	1.46E-05	1	1.46E-05	9.669145	0.0145	significant
BD	1.31E-05	1	1.31E-05	8.684428	0.0185	significant
CD	3.28E-05	1	3.28E-05	21.66088	0.0016	significant
Residual	1.21E-05	8	1.51E-06			
Cor Total	0.000187	15				

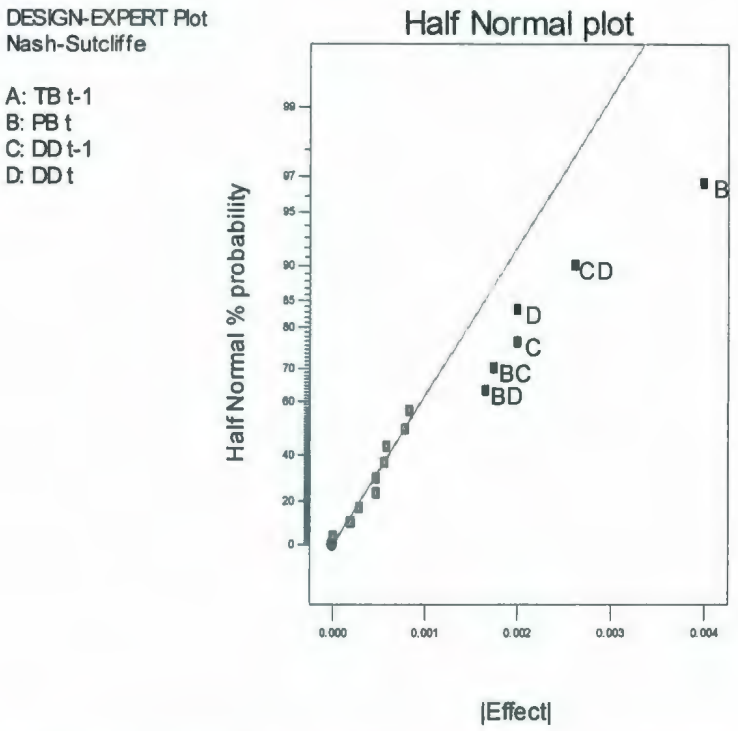


Figure 5.3 Effects plot for the 4 less important factors of BPNN at Black Brook

Table 5.6 ANOVA of the 3 more important factors of BPNN at Black Brook

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F	Judgement
Model	0.003056	4	0.000764	572.9775	0.0001	significant
A	0.002903	1	0.002903	2177.415	< 0.0001	significant
B	4.05E-05	1	4.05E-05	30.375	0.0118	significant
C	7.69E-05	1	7.69E-05	57.66	0.0047	significant
AC	3.53E-05	1	3.53E-05	26.46	0.0142	significant
Residual	4E-06	3	1.33E-06			
Cor Total	0.00306	7				

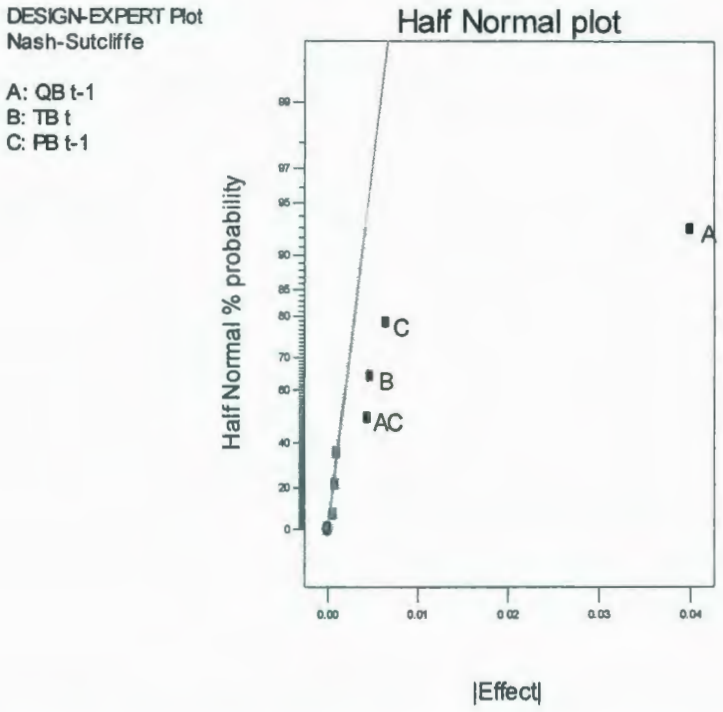


Figure 5.4 Effects plot for the 3 more important factors of BPNN at Black Brook

According to the results, the statistical significance from DOE agrees with the contributions estimated by NeuroShell2 for both BPNN and GRNN, respectively. The statistically insignificant factors which were removed from the BPNN and GRNN models are all ranked last for their contribution to the goodness of fit.

The GRNN and BPNN model are then applied using only the statistically significant input factors. After the training process, the models then applied to the validation data set to see if they can give good forecasts on the new data. The results are shown in Table 5.7

Table 5.7 Statistical results of trained ANNs for the Humber River at Black Brook

Statistical indicators	Learning set		Validation set	
Network type	GRNN	BPNN	GRNN	BPNN
Nash-Sutcliffe efficiency	0.9157	0.8861	0.8095	0.7758
r-squared	0.9164	0.8862	0.8131	0.7902
Mean squared error	94.840	128.125	158.265	186.210
Mean absolute error	5.546	6.146	6.707	7.244
Percent over 30%	24.349	22.705	17.164	17.164

The results show that both GRNN model and default setting of the BPNN model gave satisfactory performances for a 1-day ahead forecast. The GRNN model however performed a little better than the BPNN model for the Upper Humber River at Black Brook area on both learning set and validation set. Since the criterion of 'percent over 30%' (percent of outliers) is around 20%, there is still some overestimation or underestimation during the flood season. In addition, since the drainage area at Black Brook is relatively small compared with those of further downstream there is a lagging effect that cannot be avoided in the 1-day ahead forecast. This is shown in the Figures 5.5 and 5.6.

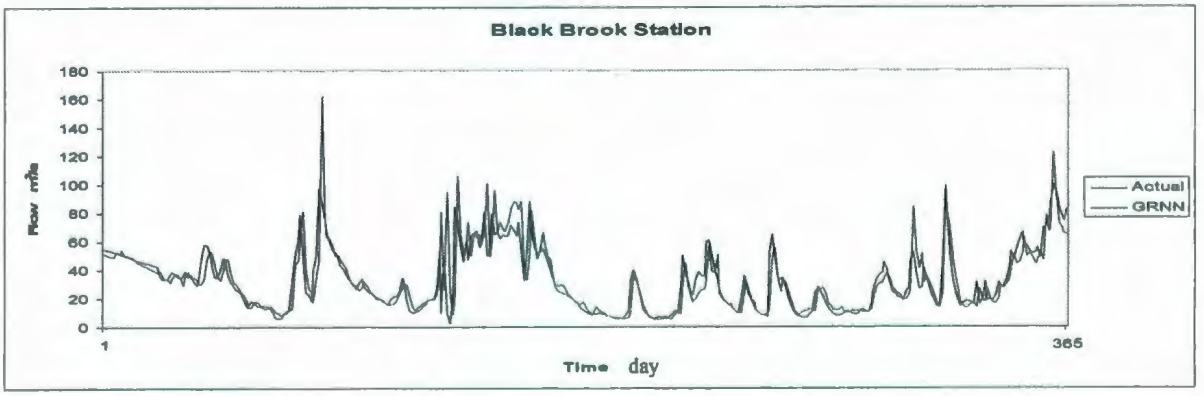


Figure 5.5 Comparison of 1-day ahead forecasts from GRNN with actual flows at Black Brook of 2007

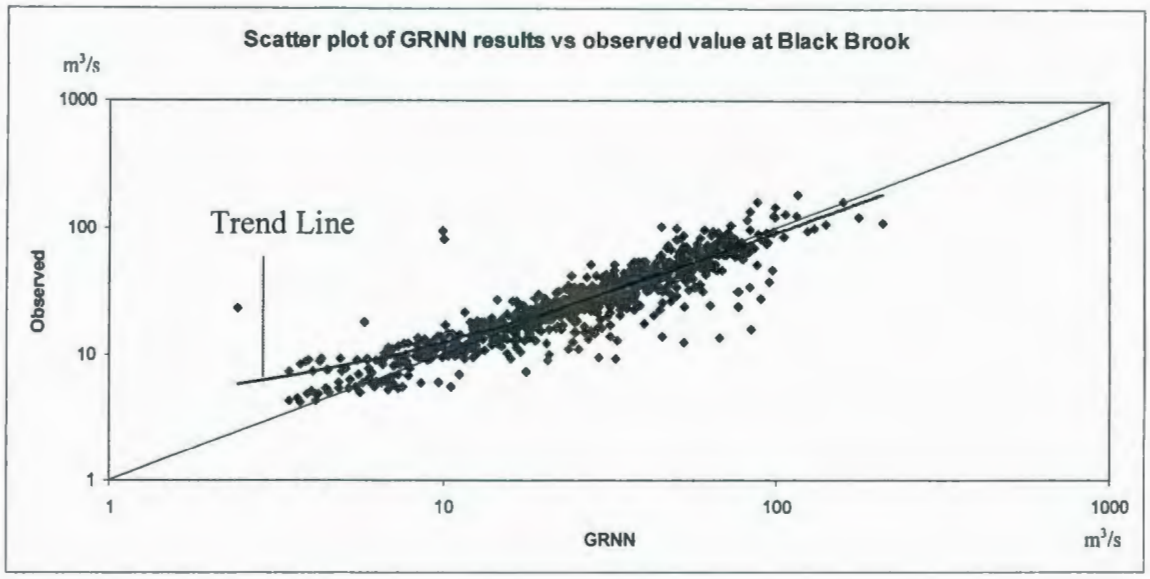


Figure 5.6 Scatter plot of GRNN results vs. observed values at Black Brook [The straight line from (1, 1) to (1000, 1000) is the line of perfect agreement.]

5.2 Modeling of the Humber River flow at Reidville

Vied from the map of Humber River Basin (Figure 1.1 in Chapter 1), the flow of the Upper Humber River near Reidville is related to its upstream river flow at Black Brook and other variables near Reidville. From the data available on hand, the temperature and precipitation measured at Adies Lake are representative of the climate of this region. The model for the flow at Reidville thus uses the climate data at Adies Lake, and the hydrometric data at Black Brook and Reidville. The potential inputs for this model thus are:

- TA_{t-1}: Air temperature at Adies Lake of 1day before
- TA_t: Air temperature at Adies Lake of current day
- PA_{t-1}: Total precipitation at Adies Lake of 1 day before
- PA_t: Total precipitation at Adies Lake of current day
- QB_{t-1}: Flow at Black Brook of 1day before

- QB_t: Flow at Black Brook of current day
- QR_{t-1}: Flow at Reidville of 1 day before
- QR_t: Flow at Reidville of current day

The output is the flow at Reidville at t+1, or the 1-day ahead forecast. The 'cumulative degree days' factor is not used because this area is not covered by heavy snow during the winter unlike the Black Brook area. The data available are from 1999 to 2008. The data from 1999 to 2002 were used for training, and data from 2003 to 2005 were used for testing. The rest of the data from 2006 to 2008 were used for validation. Both GRNN and BPNN were used to investigate which one works better for flow forecasts at this station. The calibration parameters of BPNN are set at the default values with 43 hidden neurons, 0.05 learning rate, 0.5 momentum, and 110 calibration interval.

For the Humber River at Reidville, the smoothing factors of GRNN and input strengths of BPNN are calculated as well to estimate the contribution of each variable on the network outputs.

Table 5.8 Individual smoothing factors of GRNN at Reidville

Input Variables	Individual smoothing factor	Rank
QB_{t-1}	1.10588	2
QB_t	0.17647	6
TA_{t-1}	0.00502	8
TA_t	1.00000	3
PA_{t-1}	0.05882	7
PA_t	0.44706	5
QR_{t-1}	0.65882	4
QR_t	2.92941	1

Table 5.9 Input strength of variables of BPNN at Reidville

Input Variables	Input strength	Rank
QB_{t-1}	0.08976	5
QB_t	0.05838	8
TA_{t-1}	0.06699	7
TA_t	0.11469	4
PA_{t-1}	0.13619	3
PA_t	0.08068	6
QR_{t-1}	0.14402	2
QR_t	0.24429	1

As can be seen in Tables 5.8 and 5.9, the air temperature is not as important as it is at the Black Brook area. This may be because the snowmelt which depends on temperature is not significant in this area. The flow data at Reidville are shown to be very important in both GRNN and BPNN results. Therefore, the Humber River Flow at Reidville is also highly autocorrrelated. The upstream flow at Black Brook played a bigger role in the GRNN model than in the BPNN model

GRNN Model

DOE methodology is then applied next to test for statistical significance of the input factors. As before, the less important factors are first selected. For the GRNN model, the 3 selected factors are A: TA_{t-1}, B: PA_{t-1}, and C: QB_t. From the results shown in Table 5.10 and Figure 5.7, factor B and C are significant at the 5% level with the Nash-Sutcliffe efficiency as the response, but factor A is not significant. Factor A: TA_{t-1} has the smallest smoothing factor and therefore make sense that it contributes least. Since there is a significant factor in the first step test, the rest of the factors are then assumed to be statistically significant as well. Hence, the factors used at Reidville station by the GRNN model are the following seven factors: QB_{t-1}, QB_t, TA_t, PA_{t-1}, PA_t, QR_{t-1}, and QR_t.

Table 5.10 ANOVA of the 3 less important factors of GRNN at Reidville

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F	Judgement
Model	0.000614	3	0.000205	32.81305	0.0028	significant
A	1.62E-06	1	1.62E-06	0.259876	0.6370	n.s
B	8.06E-05	1	8.06E-05	12.93684	0.0228	significant
C	0.000531	1	0.000531	85.24243	0.0008	significant
Residual	2.49E-05	4	6.23E-06			
Cor Total	0.000639	7				

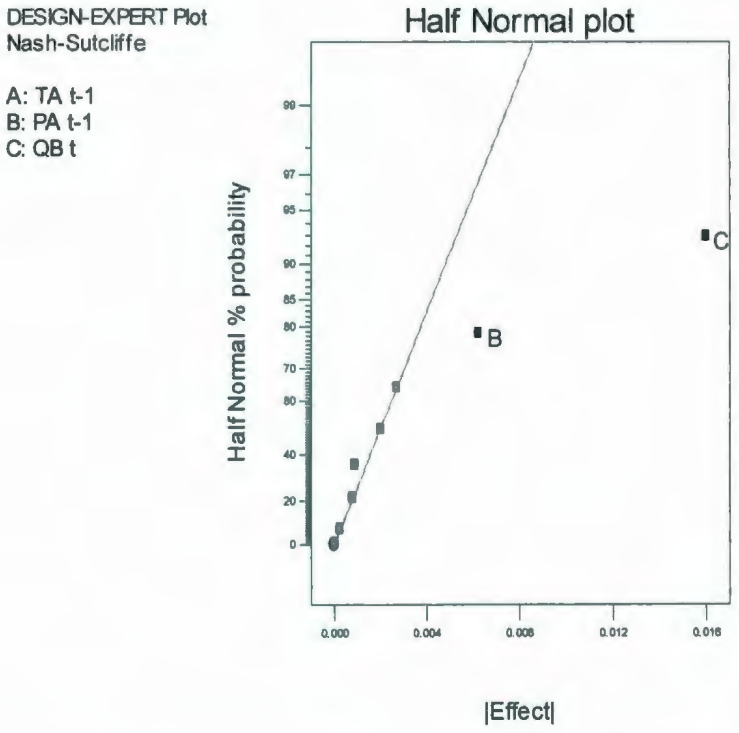


Figure 5.7 Effects plot for the 3 less important factors of GRNN at Reidville

BPNN Model

For BPNN model at this station, the selected less important factors are A: TA_{t-1}, B: PA_t, C: QB_{t-1}, and D: QB_t. From the results shown in Table 5.11 and Figure 5.8, all the factors are significant at the 5% level. This means all the factors are needed in the model. The 8 factors are all kept in the BPNN model for the Reidville station.

Table 5.11 ANOVA of the 4 less important factors of BPNN at Reidville

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F	Judgement
Model	0.001355	7	0.000194	105.7776	< 0.0001	significant
A	9.92E-06	1	9.92E-06	5.423057	0.0483	significant
B	9.22E-05	1	9.22E-05	50.36926	0.0001	significant
C	0.000355	1	0.000355	194.1985	< 0.0001	significant
D	0.00071	1	0.00071	388.166	< 0.0001	significant
AB	1.52E-05	1	1.52E-05	8.312895	0.0204	significant
BC	1.44E-05	1	1.44E-05	7.892058	0.0229	significant
CD	0.000158	1	0.000158	86.08164	< 0.0001	significant
Residual	1.46E-05	8	1.83E-06			
Cor Total	0.001369	15				

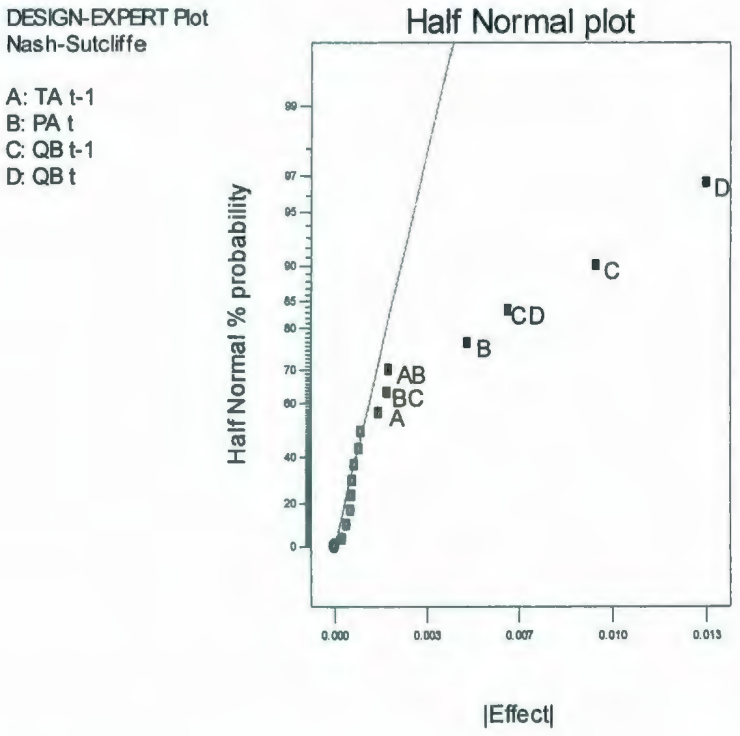


Figure 5.8 Effects plot for the 4 less important factors of BPNN at Reidville

The results from the BPNN and GRNN models for the Reidville station when applied to the learning and validation sets are shown in Table 5.12.

Table 5.12 Statistical results of trained ANNs at Reidville

Statistical indicator	Learning set		Validation set	
Network type	GRNN	BPNN	GRNN	BPNN
Nash-Sutcliffe efficiency	0.9529	0.9478	0.8836	0.8896
r-squared	0.9534	0.9480	0.8837	0.8953
Mean squared error	356.790	395.515	575.483	545.448
Mean absolute error	9.474	9.880	12.605	12.415
Percent over 30%	12.049	9.593	21.595	16.390

The results of both models at Reidville are better overall than those at Black Brook. Figures 5.9 and 5.10 also show that the lagging effect has decreased as well because the drainage area and river length of this part of the basin are larger than those at Black Brook.

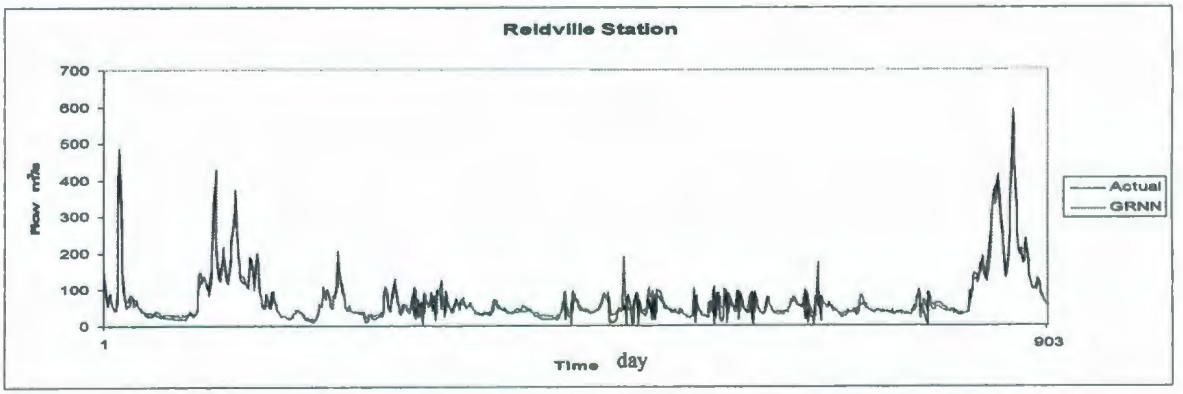


Figure 5.9 Comparison of 1-day ahead forecasts from GRNN with actual flows at Reidville from 2006 to the mid of 2008

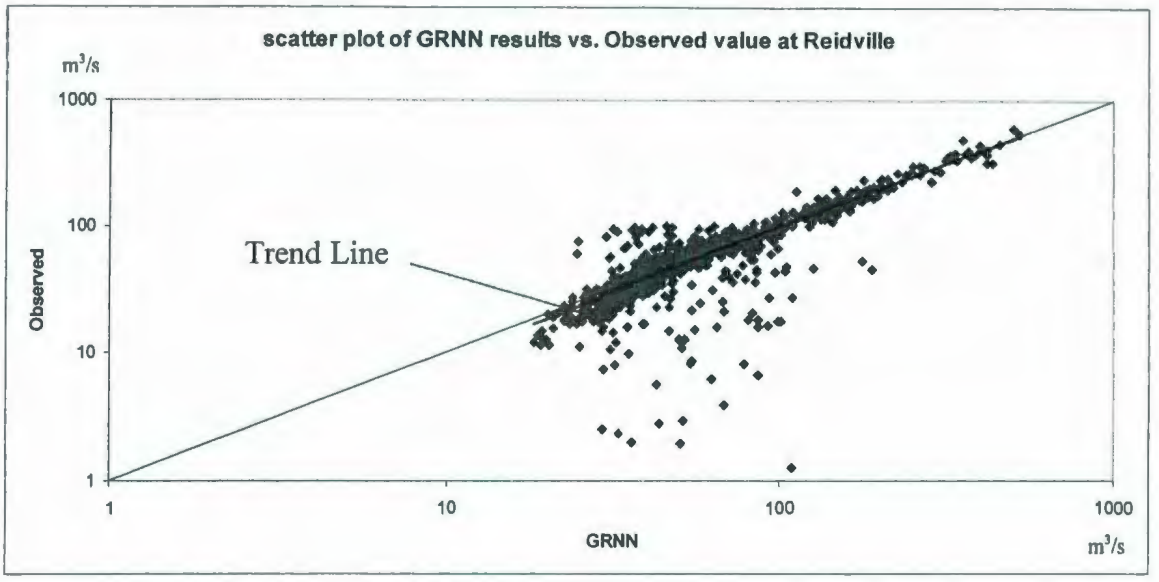


Figure 5.10 Scatter plot of GRNN results vs. observed value at Reidville. [The straight line from (1, 1) to (1000, 1000) is the line of perfect agreement.]

5.3 Modeling of the Humber River at Humber Village Bridge (Lower Humber)

From Figure 1.1, the Humber River at Humber Village Bridge monitoring site is at the outlet of Deer Lake. Thus the flow at this station is influenced by the water level of Deer Lake. There are no climate stations around this area. Climate variables are thus not considered in this model. The flow at Reidville is considered because it is the inflow into Deer Lake. Therefore the forecasted flow at Reidville may be included as in input into the flow model at Humber Village Bridge. There are thus six inputs contributing to the flow forecast at this station . They are:

- QR_{t-1}: Flow at Reidville of 1 day before
- QR_t: Flow at Reidville of current day
- WL_{t-1}: Water level of Deer Lake of 1 day before

- WL_t: Water level of Deer Lake of current day
- QV_{t-1}: Flow at Humber Village Bridge of 1 day before
- QV_t: Flow at Humber Village Bridge of current day

The output of this model is the flow at Humber Village Bridge at time t+1. The data length is the same as that at Reidville. The data from 1999 to 2002 were used for training, data from 2003 to 2005 were used for testing and the rest of the data from 2006 to 2008 were used for validation. The calibration parameters of the BPNN are similarly set at the default with 43 hidden neurons, 0.05 learning rate, 0.5 momentum, and 110 calibration interval.

For the Humber River at Village Bridge, the smoothing factors of the GRNN model and input strength of the BPNN model are also calculated to provide an estimate of the contribution of each variable on the network outputs.

Table 5.13 Individual smoothing factors of GRNN at Village Bridge

Input Variables	Individual smoothing factor	Rank
QR_{t-1}	0.01176	6
QR_t	0.41176	5
WL_{t-1}	2.78824	2
WL_t	2.23529	3
QV_{t-1}	0.64706	4
QV_t	2.94118	1

Table 5.14 Input strength of the variables of BPNN at Village Bridge

Input Variables	Input strength	Rank
QR_{t-1}	0.10380	3
QR_t	0.08100	4
WL_{t-1}	0.07785	5
WL_t	0.07501	6
QV_{t-1}	0.17916	2
QV_t	0.40245	1

For the GRNN model, the water level of Deer Lake makes almost the same contribution as the flow of one day before at Village Bridge on the network outputs. Therefore the flow at Village Bridge should be highly relevant to the water level of Deer Lake. However from the results of the BPNN model, the water level of Deer Lake is not considered important. The flow of one day before at Village Bridge has the most contribution.

GRNN Model

DOE methodology is used to estimate statistically the contribution of each factor for both GRNN and BPNN models. For the GRNN model, there are 3 factors having smoothing factors lower than 1.0. These 3 factors, A: QR_{t-1}, B: QR_t and C: QV_{t-1}will used in the first step test with the Nash-Sutcliffe coefficient as the response. The ANOVA results are shown in Table 5.15. As can be seen, factors B and C are statistically significant (P-value < 0.05) but factor A is not. Since there are factors found to be significant in the first step test, the second step test is not necessary. Therefore the factors used to develop the GRNN model for the Humber Village Bridge station is

reduced to 5. They are: QR_t, WL_{t-1}, WL_t, QV_{t-1}, and QV_t. The effects plot is shown in Figure 5.10.

Table 5.15 ANOVA for the 3 less important factors GRNN at Village Bridge

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F	Judgement
Model	0.000133	4	3.33E-05	210.752	0.0005	significant
A	6.61E-07	1	6.61E-07	4.187335	0.1332	n.s.
B	2.85E-05	1	2.85E-05	180.4828	0.0009	significant
C	8.91E-05	1	8.91E-05	564.2929	0.0002	significant
BC	1.49E-05	1	1.49E-05	94.04485	0.0023	significant
Residual	4.74E-07	3	1.58E-07			
Cor Total	0.000134	7				

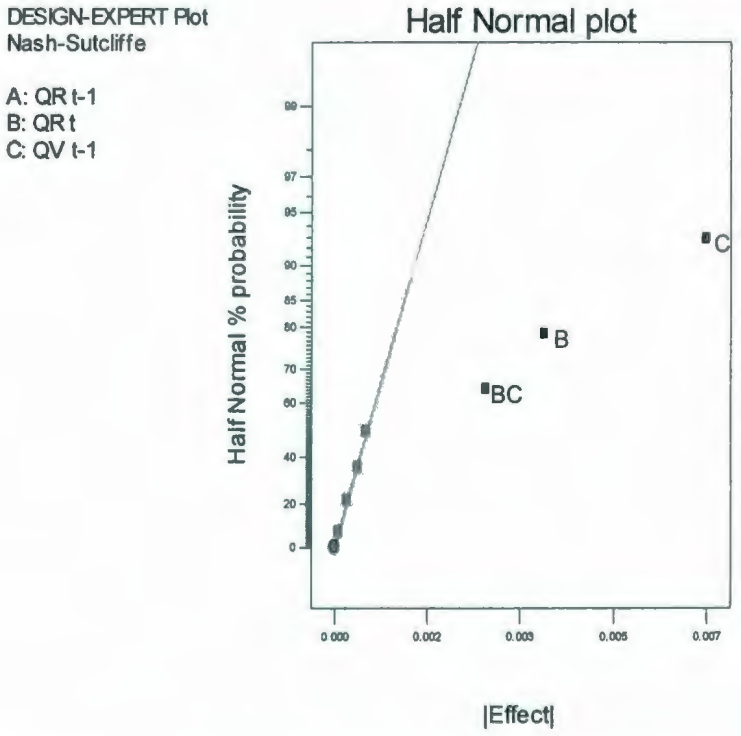


Figure 5.11 Effects plot for the 3 less important factors of GRNN at Village Bridge

BPNN Model

For the BPNN at Village Bridge station, the 3 less important factors are A: QR_t, B: WL_{t-1}, and C: WL_t. Only factor A is found to be statistically significant at the 5% level. It is not necessary to do the second step test because there is already a significant factor found in first step test. The factors finally used by the BPNN model at Village Bridge are the following 4: QR_{t-1}, QR_t, QV_t, QV_{t-1}. (see in Table 5.16 and Figure 5.12)

Table 5.16 ANOVA for the 3 less important factors of BPNN at Village Bridge

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F	Judgement
Model	6.13E-06	3	2.04E-06	816.6667	< 0.0001	significant
A	6.12E-06	1	6.12E-06	2450	< 0.0001	significant
B	0	1	0	0	1.0000	n.s.
C	0	1	0	0	1.0000	n.s.
Residual	1E-08	4	2.5E-09			
Cor Total	6.14E-06	7				

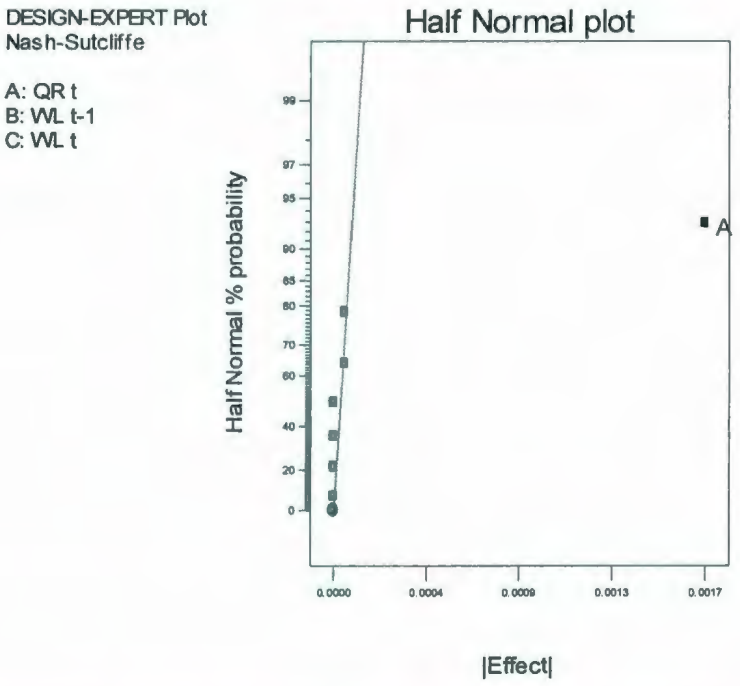


Figure 5.12 Effects plot for the 3 less important factors of BPNN at Village Bridge

The water level of Deer Lake and flow at Reidville are related to each other since the flow at Reidville is the inflow of Deer Lake. The water level of Deer Lake are somewhat decided by the flow at Reidville. In some circumstances, they could be considered as equivalent. That may be the reason that the GRNN model focused more on the water level of Deer Lake and BPNN focus more on the flows at Reidville. The BPNN and GRNN models are then trained using the statistically significant factors. The results on both learning and validation sets are given in Table 5.17.

Table 5.17 Statistical results of ANNs at Village Bridge

Statistical indicator	Learning set		Validation set	
Network type	GRNN	BPNN	GRNN	BPNN
Nash-Sutcliffe efficiency	0.9894	0.9909	0.9837	0.9901
r-squared	0.9895	0.9909	0.9838	0.9902
Mean squared error	104.805	90.148	127.415	77.486
Mean absolute error	6.052	5.455	7.249	5.544
Percent over 30%	0.058	0.029	0.111	0.111

The results of GRNN and BPNN for both learning set and validation set are excellent. The outliers are fewer than the previous two models. The Humber River at Village Bridge is the most downstream of the three stations considered. The flows do not vary as much as its upstream portion due to climatic variation. The flow is found to be "smoother" than those at Reidville and Black Brook. This can be seen in Figures 5.13 and 5.14. In addition, the drainage area is larger than that at Black Brook and Reidville. There is also almost no lagging effect of the flow forecast.

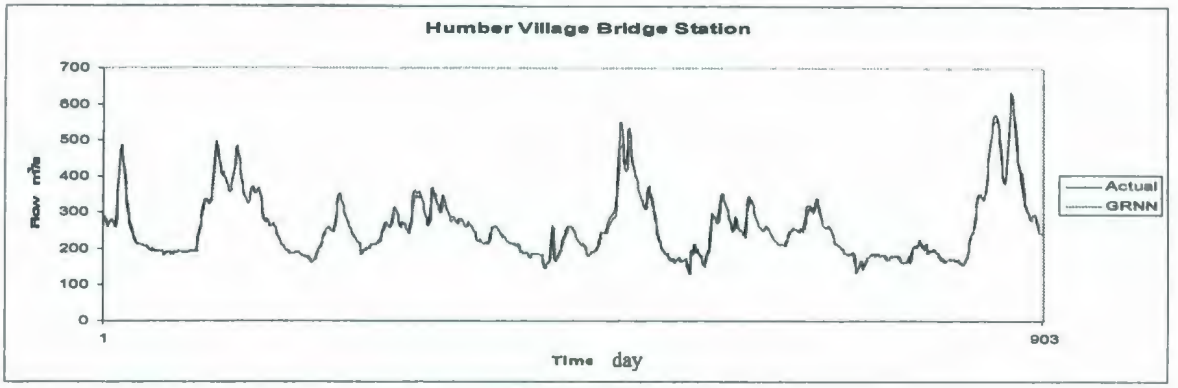


Figure 5.13 Comparison of 1-day ahead forecasts from GRNN with actual flows at Village Bridge from 2006 to the mid of 2008

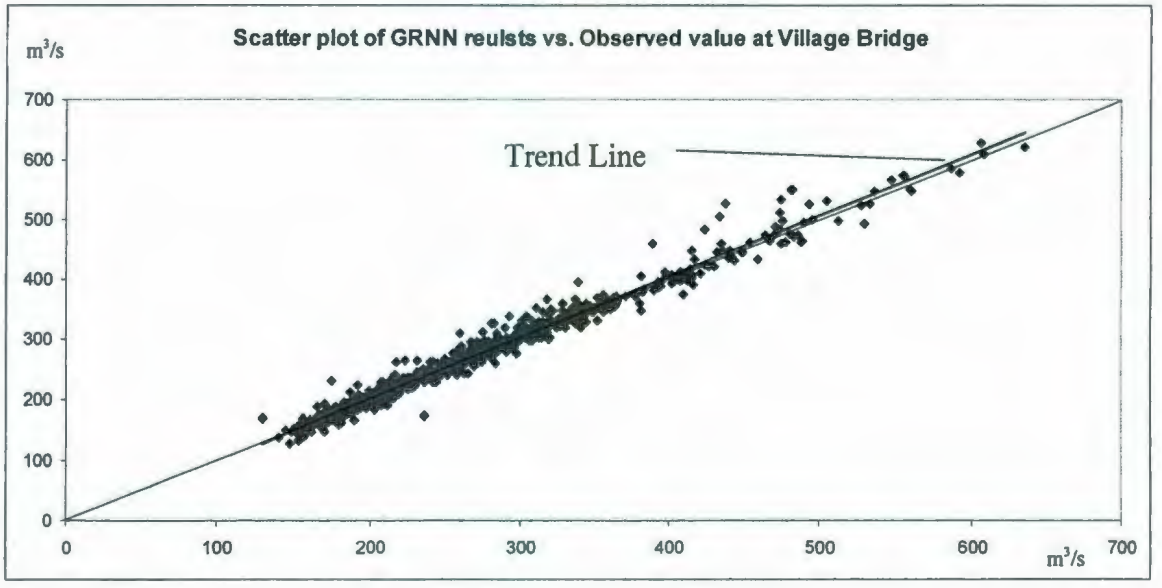


Figure 5.14 Scatter plot of GRNN results vs. observed value at Village Bridge. [The straight line from (1, 1) to (1000, 1000) is the line of perfect agreement.]

5.4 Real-time Forecasting for the 2009 Flood Season by Calibrated Models

Once the models for three different stations have been developed, they were tested at the Water Resource Management Division (WRMD) on real-time flow data during the 2009 flood season on the Humber River Basin. The performance evaluations of these ANN based models are required before a decision is made as to whether these models should be used instead of others. The real-time data supplied by the WRMD can be found at the WRMD website: (http://www.env.gov.nl.ca/wrmd/ADRS/v6/Humber/Humber_River.asp). The period of this real-time forecasting exercise started on February 25th, 2009 and ended on Jun 21st, 2009. The models produced in the previous steps then applied on the real-time data at the three stations. The results are shown in Table 5.18.

Compared with the two ANN models, the Dynamic Regression model used by WRMD also produced good forecasts for the Humber River Basin. The same indicators as ANN models are calculated in Table 5. 18 to assess performance of the Dynamic Regression model for 2009 flood season. As can be seen from the results, the ANN models are only a little better than Dynamic Regression model at Black Brook of Upper Humber. The performance of ANN models and Dynamic Regression model are practically identical at Reidiville station and Village Bridge station.

However, the routing model used by WRMD is not as good as the above three models at Black Brook and Reidville. The forecasting results of routing model for the same flood season of 2009 are compared with the other models using the same statistical indicators are shown in Table 5.18.

Table 5.18 Statistical results of 4 models at 3 stations along the Humber River (2009 Flood Season)

Statistical indicator	Black Brook				Reidville				Village Bridge			
Network type	GRNN	BPNN	DynReg	Routing	GRNN	BPNN	DynReg	Routing	GRNN	BPNN	DynReg	Routing
N-S efficiency	0.8247	0.8001	0.7946	0.4898	0.9532	0.9461	0.9617	0.5271	0.983	0.9937	0.9881	0.9658
r squared	0.8214	0.8077	0.8164	0.569	0.9614	0.9489	0.9648	0.5893	0.9833	0.9937	0.9883	0.9762
Mean squared error	259.751	281.638	247.868	775.483	320.266	506.091	353.599	5255.79	84.1771	226.753	166.202	350.22
Mean absolute error	8.9619	11.3531	10.2155	14.1126	18.2073	14.5505	11.7731	56.5879	6.1204	10.6173	8.6717	9.6164
Percent over 30%	0.1066	0.1207	0.0909	0.3967	0.0431	0.0259	0	0.7903	0	0	0.0069	0.0244

The results at the Humber River at Black Brook are not as good as those of the other two stations while the doing validation. The decreased performance at Black Brook was because the model did not cover the whole year of data. It just modeled the flood season which is the hardest part for forecasting. Although the results are a little worse than expected, they are still quite satisfactory by real-time flow forecasting standards. Figure 5.15 shows the predicted flows compared with the actual flows at the three stations.

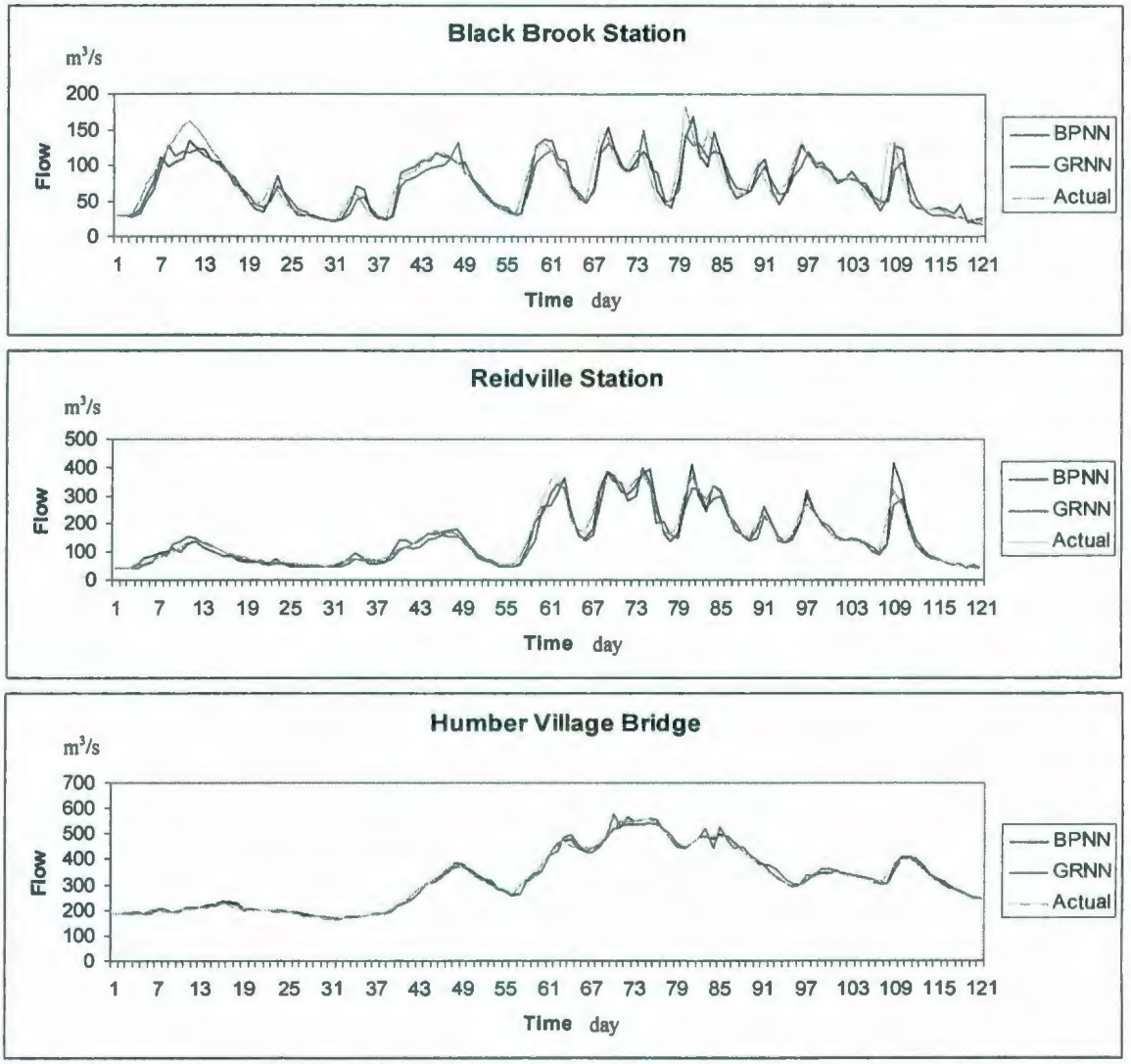


Figure 5.15 Comparison of the forecasts from BPNN and GRNN and actual flows at Black Brook, Reidville, and Village Bridge.

5.5 Discussion

From the results of GRNN and BPNN for all the 3 stations along Humber River, GRNN gave slightly better results than BPNN at the Black Brook of Upper Humber River. This may be because the GRNN is better at dealing with the cumulative degree days as a snow melt factor than BPNN. For the two stations of the Lower Humber, the results of GRNN and BPNN are practically same. Both of them are good at flow forecasting for the non-snow area.

Beside the performance of the models, the two ANN models have totally different algorithm and training methods. BPNN is based on the activation function (i.e. sigmoid function). The inputs and outputs are connected by several such activation functions. The training is based on the weight change according to the backpropagation of errors. On the other hand, the GRNN algorithm is based on the comparing of distance of input between new patterns and old patterns on each dimension to estimate the distance of output between new patterns and old patterns. The training of GRNN is then to estimate the relationship between input distance and output distance on every two given patterns. Although the DOE results agree with the contributions estimated by NeuroShell2 for both BPNN and GRNN, BPNN is more mathematically correct and makes more sense in terms of hydrology. This is because in BPNN, all the inputs are scaled from 0 to 1 so that the weights are comparable, while it is not done in GRNN.

The use of DOE methodology provides a statistical basis for the extracting of factors to be used in an ANN. Although NeuroShell2 can provide the contribution list after the model has been trained it is still not clear whether the input factors should be included in the model or not. DOE

solves this problem statistically. If an input factor is statistically significant by DOE testing, it will be included in the model, otherwise it can be left out.

Chapter 6

Conclusions and Recommendations

6.1 Conclusions

Artificial neural network (ANN) methodology was used in this thesis to develop models to produce 1-day ahead forecasts for the Upper and Lower Humber River Basin. The backpropagation algorithm was first applied and its several parameters were calibrated by DOE methodology. The calibration of these parameters was to seek the most appropriate combination that can optimize the model. However, the results of the calibration exercise showed that the choice of parameter values had little effect on the model performance. Therefore, the default settings of BPNN parameters were suggested for further work. It was also found that the cumulative degree days is an important factor for streamflow forecasting in the heavily snow covered areas above Black Brook. To develop a model with relevant input variables, the hydrometereologic factors used were tested for statistical significance by DOE methodology at a significance level of 5%. This approach provides an objective test to select variables in an ANN model. The input factors used at each station by BPNN and GRNN are shown in Table 6.1. The results showed that both GRNN and BPNN models provided much better forecasts than that of the routing based model but were only slightly better in some cases than the Dynamic Regression model developed by Picco (1996) and used by the WRMD.

Table 6.1 Input factors used by the BPNN and GRNN models

Black Brook		Reidville		Village Bridge	
GRNN	BPNN	GRNN	BPNN	GRNN	BPNN
QB_{t-1}	QB_{t-1}	QB_{t-1}	QB_{t-1}	QR_t	QR_{t-1}
TB_{t-1}	TB_{t-1}	QB_t	QB_t	WL_{t-1}	QR_t
TB_t	TB_t	TA_t	TA_{t-1}	WL_t	QV_t
PB_t	PB_t	PA_{t-1}	TA_t	QV_{t-1}	QV_{t-1}
DD_t	DD_{t-1}	PA_t	PA_{t-1}	QV_t	
QB_t	DD_t	QR_{t-1}	PA_t		
	QB_t	QR_t	QR_{t-1}		
			QR_t		

The BPNN model has several parameters to calibrate, but once the parameter combinations are set, the training process only takes ten to twenty minutes. On the other hand, the GRNN model does not have parameters to calibrate, but it is more time consuming during training especially for the large training patterns. The GRNN model can take up to three to four hours for training. Although the results of all three stations are satisfactory, the performance of the two stations at the Lower Humber is much better than the one at the Upper Humber. Several reasons that may have caused the difference in performance at the three stations are:

1. The Upper Humber Station at Black Brook is highly influenced by snowmelt which is not considered at the other two stations. The cumulative degree day index, while a statistically significant factor, may not be able to capture the snowmelt component accurately.

2. The Upper Humber drainage area at Black Brook is much smaller than the other two stations. For a small basin, using daily flows may not be a short enough time frame to respond to the input changes. The flow may change rapidly with a sudden change in one or more input factors. Perhaps the more appropriate data to use are hourly or even half-day data. This is known to produce better performance on small river basins. However, these data are not conveniently available for this study.
3. The Upper Humber River flows through more mountainous area than the lower portion. Some physiographical parameters which were not considered in the model may have affected the accuracy of the model as well. On the other hand, at the Lower Humber around Deer Lake, the flow is through the plains. The flows are thus less affected by the physiographical conditions.

The use of DOE methodology was shown to be an efficient model calibration strategy. It can provide inexperienced users an easy way to deal with a new model. Although DOE also requires some time for calibrating, it provides a more systematic process for model calibration than the traditionally used trial and error or changing one parameter at a time method. The use of DOE methodology is not restricted only to the calibration of model parameters but it can also be used to provide an objective statistical approach to extracting the input hydrometerologic factors that best contribute to the goodness of fit of the ANN models. Although the ANN software, NeuroShell2, provides a measure of the contribution of the input factors, no statistical measure is attached with it; hence there is no guarantee that a particular input factor is actually statistically significant.

One puzzling aspect of using ANN models is that the input factors that are considered to be statistically significant are different depending on which training algorithm is used. That is, BPNN and GRNN gave a different set of input factors and it is difficult to assess which model actually is more realistic from a hydrology point of view. The weights of the BPNN model are actually more comparable as all inputs have been scaled from 0 to 1.

6.2 Recommendations

In this thesis, only 1-day ahead forecasts are provided. This is because the 1-day ahead forecast provides the most accurate result for small watersheds, and it is usually sufficient. When it is necessary to provide 2-day ahead forecasts, the forecasts can be processed in two steps. First, 1-day ahead forecasts are generated. Then, the results of the 1-day ahead forecasts with other forecasted meteorological data are used as inputs for the 2-day ahead forecasts using the same model as 1-day ahead forecasts. The results of the two step model do not usually perform very well because it is subject to more errors due to the forecasted input factors from several sources. But, that is the only way to provide more than the 1-day ahead forecast. Since the Humber River Basin is a small basin, the hourly or half-day ahead forecast may provide better forecasts than 1-day ahead. But the hourly data provided by WRMD is not complete. In order to make the forecasts more accurate, half-day data or hourly data need to be recorded if possible.

The input factor of cumulative degree-days used in this study is a simple formulation based on temperatures above a threshold. Hence it is only a proxy variable to measure the snowmelt amount. In actuality, there are many issues and complexities in modeling the accumulation and melting of snow packs on a complex topography. Energy budget and snow pack evolution

models may be needed for a detailed analysis. Even with the degree-day method, there are also many forms can be used for different conditions. Further analysis of the snow accumulation and melting problem for the Upper Humber Area may improve the performance of forecasts together with the use of shorter duration flow data.

In addition, the physical reasonableness of the ANN models needs further investigation. It would be of interest to find out why the GRNN and BPNN models use a different set of input factors to give practically the same goodness of fit. For example, Q_R and W_L were used by BPNN and GRNN representatively. Probably Q_R and W_L are almost exactly the same if the relationship between Q_R and W_L are taken into account.

From the performance of the 4 models used for the 2009 flood season forecasts, the ANN and dynamic regression models should both be used for future forecasts to provide a check to each other since they provided identical practically the same performance especially in the less-snow covered area. The cumulative degree day index is recommended for inclusion as a variable in the dynamic regression model so that the snowmelt can be represented in the dynamic regression model. It is further suggested that both ANN and dynamic regression be used together for the next few years so that both models can be further calibrated and validated with more data.

References

Abrahart, R. J. 2005. Neural Network Modelling: Basic tools and Broader Issues. Neural Networks for Hydrological Modeling. pp 15-37

Beven. K. 2001. Rainfall-Runoff Modeling: A Primer. John Wiley and Sons Ltd.

Baxter, C. W., Smith, D. W., and Stanley, S. J. 2004. A Comparison of Artificial Neural Networks and Multiple Regression Methods for the Analysis of Pilot-Scale Data. J. Environmental Engineering and Science. Vol. 3(S1), pp. S45-S58

Box, G.E.P. and Jenkins, G.M. 1976. Time Series Analysis: Forecasting and Control. Wiley Series in Probability and Statistics

Campolo, M., Soldati, A., and Andreussi, P. 2003. Artificial Neural Network Approach to flood Forecasting in the River Arno. Hydrological Sciences Journal/ Journal des Sciences Hydrologiques, Vol. 48(3), pp. 381-398

Cornell, J. A. 1990. "How to Apply Response Surface Methodology", Volume 8, American Society for Quality Control, USA, 82 pages.

Cumming Cockburn and Associates. 1984. Hydrotechnical Study of the Steady Brook Area, Main Report, Canada-Newfoundland Flood Damage Reduction Program, St. John's, NF

Danh, N. T., Phien, H. N., Gupta, A. D. 1999. Neural network models for river flow forecasting. Asian Institute of Technology, Water SA, Vol. 25, No. 1, pp. 33-40

Daniell, T. M. 1991. Neural Networks – Applications in Hydrology and Water Resources Engineering. International Hydrology & Water Resources Symposium Perth. pp. 797-802

Dawson, C. W., Abrahart, R. J., Shamseldin, A. Y., and Wilby, R. L. 2006. Flood Estimation at Ungauged Sites Using Artificial Neural Networks. Journal of Hydrology Vol. 319, pp. 391-409

Fleming, S. W., and Quilty, E. J. 2007. Toward a Practical Method for Setting Screening-Level, Ecological Risk-Based Water Temperature Criteria and Monitoring Compliance. Environmental Monitoring and Assessment. Vol. 131, pp. 83-94

Flood, I., and Kartam, Nabil. 1994. Neural Networks in Civil Engineering I: Principles and Understanding. Journal of Computing in Civil Engineering, Vol. 8, No. 2, pp. 131-147

Goodrich Robert L. 1989. Applied Statistical Forecasting. Business Forecast Systems Inc. Belmont, MA

Gourrion, J. 2000. Ku-band wind speed model functions via neural network methods. Technical report DOS, French Research Institute for Eploitation of the Sea, vol.2000-02,

Jain, P., and Deo, M. C. 2006. Neural Networks in Ocean Engineering. Ships and Offshore Structures Vol. 1, No. 1, pp. 25-35

Johnson, N. L. and Leone, F. C. 1977. "Statistics and Experimental Design in Engineering and the Physical Sciences", Volume II, John Wiley & Sons, New York, 490 p.

Kerh, T. F., and Lee, C. S. 2006. Neural networks forecasting of flood discharge at an unmeasured station using river upstream information. Advances in Engineering Sofrware, No. 37, pp. 533-543

Kneale, P. E. K., See, L. M., and Abrahart, R. J. 2005. Why Use Neural Networks? School of Geography, University of Nottingham, University of Leeds, UK

Li, X., Smith, D. W., and Prepas, E. E. 2008. Artificial Neural Network Modelling of Nitrogen in Streams: with Emphasis on Accessible Databases. Proceedings of the CSCE 2008 Annual Conference Quebec.

Myers, R. H. and Montgomery, D. C. 1995. Response Surface Methodology: Process and Product Optimisation Using Designed Experiments, John Wiley & Sons, Inc. 700 p.

Pankratz, Alan. 1991. Forecasting with Dynamic Regression Models. John Wiley and Sons, New York,

Peeters, A. G. 1998. Cumulative temperatures for prediction of the beginning of ash (Fraxinus excelsior L.) pollen season. Aerobiologial. Vol. 14, pp. 375-381.

Picco, R. 1996. Flow Forecasting for the Humber River Basin Spring 1996. Water Resources Management Division Department of Environment and Labour, Government of Newfoundland and Labrador.

Rolling, K. 2008. Personal Communications.

Savelieva, E. 2004. Automatic Spatial Prediction with General Regression Neural Network (GRNN). Nuclear Safety Institute (IBRAE) of Russian Academy of Science. Vol. 1, No. 2

Singh. 1989. Hydrologic Systems: Watershed Modeling Volume II. Prentice-Hall, Inc. Division of Simon & Schuster Englewood Cliffs, New Jersey 07632

Statistic Canada. 2006 Community Profile – Corner Brook, Newfoundland and Labrador

Sulistiyono, H. 1999. Rainfall- Runoff Model Calibration using Experimental Designs and Response Surface Methodology. MEng thesis, Faculty of Engineering and Applied Science Memorial University of Newfoundland

Suzuki, R., Nomaki, T., and Yasunari, T. 2003. West-east contrast of phenology and climate in northern Asia revealed using a remotely sensed vegetation index. International Journal of Biometeorology. Vol. 47, No. 3, pp. 126-138

Ward System Group, Inc. 2000. "Neuroshell 2, Release 4.0"